How Should Linking Accuracy Be Measured and Across-Crime Similarity Assessed in Behavioural Linkage Analysis?

by

Holly Ellingwood

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirement for the degree of

Master of Arts

in

Psychology

Carleton University
Ottawa, ON

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-93557-6

Our file Notre référence
ISBN: 978-0-494-93557-6

Abstract

The ability to link serial crimes accurately by relying on crime scene behaviours may be affected by the measure used to assess linking accuracy and by the coefficient used to assess across-crime similarity. To examine these issues, the present study first manipulated two factors - the variability of across-crime similarity scores used to make linking decisions and variability in the base rate of linked crimes - to determine the robustness of several measures of linking accuracy - the correlation coefficient (*r*), the area under the curve (*AUC*), and the odds ratio (*OR*) - to these manipulations. Both *r* and the *AUC* were sensitive to restriction of range in the predictor variable, however the *AUC* was more affected. The *AUC* was robust to base rate variability while *r* was not. Both *r* and the *AUC* could easily be interpreted, whereas problems were encountered with the *OR*. Based on these findings, both *r* and the *AUC* were then used to assess the level of linking accuracy that could be achieved when using one of three similarity coefficients - Jaccard's coefficient (*J*), the simple matching index (*S*), and the Sorensen-Dice index (*SD*) – to measure across-crime similarity. Based on analyses using both serial burglary and serial rape data, the results demonstrated that, regardless of what measure of linking accuracy was used, no significant differences were found between the different similarity coefficients except in the case of *S*, in a single instance when it performed significantly worse than *J* and *SD*.

Acknowledgements

This research project could not have been accomplished without the assistance of various people to whom I would like to express my deepest gratitude. First and foremost, I would like to thank my advisor Dr. Craig Bennell. His steadfast guidance and insight along with his support, despite the demands of a heavy schedule on his part, were appreciated beyond what mere words of thanks can express and likely do not suffice. I am most grateful to Tamara Melnyk, who, by offering her time and experience in crime linkage research, helped me immeasurably. The quality of this final product would not have been possible without her so generously devoting her time and sharing her expertise, and I am grateful for that as well as her support and friendship throughout this endeavour. I would also like to thank my Committee Members, Dr. Kevin Nunes and Dr. Evelyn Maeder, whose thoughtful comments helped make this a stronger research enterprise and I thank them for giving me their invaluable time and consideration. The good advice, support, and friendship from my graduate colleagues, Karla Emeno, Rebecca Mugford, and Deanna Whelan, were invaluable on both an academic and personal level for which I am extremely grateful. My abiding gratitude also goes to Etelle Bourassa, Graduate Studies Administrator for the Department of Psychology at Carleton University, who was always available to answer any administrative concerns or otherwise, that I had throughout this endeavour, making a difficult undertaking that much less arduous thanks to her support and availability. Finally, to my mother, Lawanda Willar, I express the utmost gratitude for her unequivocal support and encouragement. The final product is a reflection of my hard work and the many hurdles that had to be overcome, and would not have been possible were it not for the people mentioned above to which I am indebted.

Table of Contents

List of Tables

List of Figures

List of Appendices

How Should Linking Accuracy Be Measured and Across-Crime Similarity Assessed in Behavioural Linkage Analysis?

When investigating crimes, one of the challenges police investigators face is the task of correctly deciding whether multiple crimes have been committed by a single offender (i.e., serial crime; Grubin, Kelly, & Brunsdon, 2001). Serial crime is defined by the Federal Bureau of Investigation (F.B.I.) as at least three of the same type of crime committed by the same offender in separate events (Douglas, Burgess, Burgess, & Ressler, 1992; Kocsis & Irwin, 1998; Ressler, Burgess, & Douglas, 1988). Attempting to conclude whether a series of the same type of crime has been committed by a serial offender is particularly problematic to accomplish when physical evidence, such as DNA, cannot be found, or when eyewitness reports are lacking (Woodhams, Hollin, & Bull, 2007). By correctly linking separate crimes to form a crime series, a more efficient police investigation can take place (Grubin et al., 2001; Labuschagne, 2010; Santtila, Junkkila, & Sandnabba, 2005). In contrast, incorrectly linking crimes that are not committed by the same offender would have a negative impact on any investigation, potentially leading investigators on a wild goose chase.

In cases where there is a lack of physical evidence, investigators must use behavioural information to establish any crime linkages (Woodhams et al., 2007). Of course, such behaviours cannot be observed directly. Therefore, police investigators either have to rely on clues left at crimes scenes to determine how the crimes were committed by the offender, or they must rely on the victims/witnesses of crimes to tell them how the offender behaved. Identifying behavioural patterns across crime scenes in order to establish if the crimes under investigation are committed by the same offender is

generally accomplished using an investigative technique known as behavioural linkage analysis (BLA).

BLA is used by crime analysts and specially trained police officers all over the world including Canada, the United States, the United Kingdom, Australia, New Zealand, the Netherlands, Belgium, Switzerland, France, Germany, and the Czech Republic (Royal Canadian Mounted Police; RCMP, n.d.). BLA is also increasingly being presented as evidence in court when questions are raised about whether a defendant may be responsible for multiple crimes (Markson, Woodhams, & Bond, 2010; Woodhams & Labuschagne, 2011). Examples of police units that use BLA on a regular basis include the Investigative Psychology Section of the South African police service and the Serious Crime Analysis Section in the United Kingdom (Woodhams & Labuschagne, 2011).

Two key assumptions must be established in order for BLA to actually be effective (Canter, 1995). Offenders must behave in a relatively similar fashion across the crimes they commit (behavioural stability), and they must exhibit behaviours that are different from those exhibited by other offenders who are committing similar types of crimes (behavioural distinctiveness). A number of studies over the past few decades have shown that a reasonable degree of behavioural stability and distinctiveness is demonstrated by serial offenders, making it possible to successfully link serial crimes using crime scene behaviours. This has been found in cases of serial burglary (Bennell & Canter, 2002; Bennell & Jones, 2005; Goodwill & Alison, 2006; Markson et al., 2010; Melnyk, Bennell, Gauthier, & Gauthier, 2011); serial robbery (Woodhams & Toye, 2007), serial car theft (Tonkin, Grant, & Bond, 2008), serial homicide (Melnyk et al., 2011; Santtila et al., 2008), serial arson (Ellingwood, Mugford, Melnyk, & Bennell, in

press; Santtila, Fritzon, & Tamelander, 2004), and serial sexual assault (Bennell, Gauthier, Gauthier, Melnyk, & Musolino, 2010; Bennell, Jones, & Melnyk, 2009; Grubin et al., 2001; Häkkänen, Lindlöf, & Santtila, 2004; Mokros & Alison, 2002; Santtila et al., 2005; Woodhams, Grant, & Price, 2007; Woodhams & Labuschagne, 2011; Yokota, Fujita, Watanabe, Yoshimoto, & Wachi, 2007).

These sorts of results are typically treated as evidence that offender behaviour is determined to a large extent by internal predispositions to behave in particular ways (e.g., Greene, 1989; Meyer, 1990). Indeed, much like personality psychologists do (e.g., Funder & Colvin, 1991; Furr & Funder, 2003), it is assumed that offenders possess certain traits, or at the very least, preferred offending styles (Canter, Bennell, Alison, & Reddy, 2003), that lead to stable individual differences in the way they commit their crimes over time. This is presumably what allows crimes to be linked to a common offender on the basis of one's crime scene behaviour and crimes committed by different offenders to be differentiated from one another.

With that being said, no research has ever found that offenders are so stable that perfect linking accuracy can be observed. Indeed, despite what fictional accounts might say, crime scene behaviours, like non-criminal behaviours, are not always displayed consistently across situations (Bennell & Canter, 2002). This is because nearly all behaviours, including criminal behaviours, are influenced by situational factors (Cervone & Shoda, 1999; Mischel, 1968). For example, in the criminal domain, it is now accepted that situational, learning, and maturational factors can impact the crime behaviours of serial offenders (Douglas & Munn, 1992), thus decreasing our ability to successfully link crimes.

The current study focused on two issues that are important when considering the use of BLA. The first issue that was examined in this thesis relates to how linking accuracy should actually be measured and whether certain measures of accuracy are more robust than others across conditions that vary naturally in the linkage context (i.e., the variability of the predictor variable, usually across-crime similarity scores, and the base rate of linked crimes). The second issue that was examined relates to the similarity measure that should be used to assess the level of across-crime behavioural similarity/distinctiveness when carrying out BLA. Each of these issues will be discussed in more detail within the sections that follow.

How Should Linking Accuracy Be Measured?

BLA is essentially a diagnostic task, much like other diagnostic tasks that are seen in fields as diverse as radiology (predicting cancer) and meteorology (predicting storms). The goal with BLA is to accurately determine when a set of crimes has been committed by the same offender and when the crimes have been committed by different offenders (Bennell & Canter, 2002). In diagnostic terms, the task therefore is to distinguish signals (linked crimes) from background noise (unlinked crimes). In every type of diagnostic task, one of the central challenges is to determine how accurately the diagnostic decisions are made. In other words, in the context of BLA, how good are we at actually establishing accurate crime linkages?

When answering this question researchers are typically interested in assessing the accuracy of some sort of diagnostic decision making procedure. While the performance of human decision makers can and has been assessed (e.g., Bennell, Bloomfield, Snook, Taylor, & Barnes, 2010), researchers have become increasingly interested in assessing

how statistical predictions rules (SPRs) perform on the linking task. Presumably this is due to the fact that SPRs have been shown, historically, to outperform human judges on tasks like BLA (Grove & Meehl, 1996). A number of SPRs have been proposed in this context (e.g., Bennell & Canter, 2002; Grubin et al., 2001; Santtila et al., 2004; Tonkin et al., 2008). Typically these tools take into account the level of behavioural similarity that exists across two crimes and produce as output the likelihood that the crimes are actually committed by the same offender (Bennell & Canter, 2002).

There are a number of different ways in which the accuracy of such procedures can be determined. For example, on the basis of the predicted likelihoods/probabilities/estimates produced by a SPR (or some other procedure), one can select a particular cut-off to use for making "linked" and "unlinked" decisions, and then determine the percentage of correct and incorrect decisions that are made when using that cut-off. This is very similar to the procedure adopted by Grubin et al. (2001) who applied an algorithm to a set of crimes, and then examined the 10% of crimes that were most similar to target offences to assess how many of those crimes were actually linked to the target offence.

While this approach certainly allows the user to assess the accuracy of their procedure for making linking decisions, this approach has several disadvantages. For example, it is unclear how one should go about selecting an appropriate cut-off. Who is to say, for example, that the 10% cut-off adopted by Grubin et al. is a sensible cut-off? Compounding this problem further is the fact that linking accuracy will vary as a function of the cut-off chosen because decision outcomes (e.g., whether a "linked" decision is correct [a hit] or incorrect [a false alarm] will fluctuate as cut-offs are made more lenient

[e.g., examining the top 50% of similar crimes] or more strict [e.g., examining the top 1% of similar crimes]). Given this, it would be difficult when using an accuracy measure based on the percentage of crimes correctly classified at a given cut-off to determine the predictive accuracy associated with a decision making approach (we would only know how the approach performs when using a specific threshold for making decisions).

Over the past decade, one of the most common methods for assessing linking accuracy (and for bypassing the problems mentioned above) has been to rely on the area under the curve (*AUC*), which is associated with receiver operating characteristic (ROC) analysis (Bennell & Canter, 2002). This method has also become the norm in other fields (Swets, 1988). This curve represents the ratio of hits to false alarms that can be made on any two-alternative, yes/no type discriminatory task, such as BLA, across the various thresholds that can be used to determine when a positive decision should be made (that crimes are "linked" in the present case). The decision thresholds in the case of SPRs typically relate to various cut-offs along the probability values produced by the tool when applied to a set of crimes. As indicated above, hits refer to decisions that crimes are linked when they are, while false alarms refer to decisions that crimes are linked when they are not.[1] The value of the *AUC* ranges from 1.00, indicating perfect diagnostic accuracy, to 0.00, indicating perfect inaccuracy, with values of 0.50 representing chance accuracy.

As indicated above, the *AUC* has become one of the most common methods for assessing accuracy in BLA and it is now reported in most studies. For example, Markson

[1] The other two types of decision outcomes that are possible when conducting BLA, misses and correct rejections, are also modelled on a ROC graph. They are typically ignored, however, because they are the complements of hits and false alarms and therefore, we automatically know the miss rate and correct rejection rate once we know the hit rate and false alarm rate. Specifically, misses = 1 – hits and correct rejections = 1- false alarms.

et al. (2010) studied a sample of 160 serial residential burglaries committed in the UK (*N* = 80 offenders). They used *AUC*s to evaluate the use of logistic regression models applied to a sample of linked and unlinked crimes. Results indicated that high levels of predictive accuracy were achieved when using geographic proximities (how close the crimes were to each other) and temporal proximities (how close in time the crimes were to each other) combined (*AUC* = .95). However, much lower levels of accuracy were obtained when using measures of behavioural similarity based on traditional MO indicators like entry behaviours (*AUC* = .54) and property stolen (*AUC* = .58).

In another example of the use of the *AUC* in a BLA study, Tonkin et al. (2008) were able to use logistic regression analysis to accurately discriminate linked from unlinked crimes on the basis of crime scene behaviours for 386 car thefts committed by 193 offenders. The ROC analysis indicated that the distance between car dump sites (*AUC* = .77) and car pickup sites (*AUC* = .81) were particularly powerful predictors of linked and unlinked car thefts. This was not the case for variables such as target selection (type of property selected; *AUC* = .57) and target acquisition (method of stealing vehicle; *AUC* = .56).

There are a number of advantages associated with the *AUC* as a measure of linking accuracy, which explains why it has become so common. First, unlike traditional measures of accuracy, such as the percent correctly classified, the *AUC* is independent of any decision threshold (Bennell & Jones, 2005). In other words, the ROC curve, which is used to calculate the *AUC*, covers the full range of thresholds that can be used when making linking decisions (hit values and false alarm values can be calculated and plotted for each and every threshold). The *AUC* therefore has the advantage of not being biased

by threshold placement and thus, it is a more valid measure of accuracy (i.e., it allows us to determine the true discriminatory power of a decision making tool or procedure rather than the accuracy of that tool or procedure under the limited conditions of a single threshold).

A second advantage to using the *AUC* as a measure of linking accuracy for BLA is that the measure is not expected to be impacted by the base rates in the sample of cases to which the decision making tool is applied. This is because the *AUC* is based on the proportion of decision outcomes rather than their frequencies (Swets, 1988). In the case of BLA, the base rate refers to the proportion of crimes in a sample of crimes pairs that are actually committed by the same offender (and thus should be linked). The fact that the *AUC* is unaffected by base rates is particularly helpful because it allows researchers to compare levels of linking accuracy that are achieved across different samples of crimes that may vary with respect to base rate. This will be especially true if the samples in question relate to different crime types, as will be the case in the current study (see the discussion of Phase 2 below).[2]

Despite these advantages, however, a recent study that examined how the *AUC* performs relative to other accuracy measures in the context of risk assessment decisions (specifically the correlation coefficient (*r*) and the odds ratio (*OR*) from logistic regression) found that the *AUC* may have limitations in assessing diagnostic accuracy. Specifically, Hanson (2008) examined the predictive accuracy of *r*, the *AUC*, and the *OR*

[2] For example, the base rate of linked crimes in a sample of serial burglaries would be expected to be much higher than the base rate of linked crimes in a sample of serial homicides or rapes. It is not uncommon to read about average series lengths in serial burglary cases reaching 20 crimes (e.g., Snook, Zito, Bennell, & Taylor, 2005) and series exceeding 50+ crimes are certainly not uncommon (e.g., Wright, Decker, Redfern, & Smith, 1992). These sorts of very long, linked crime series are much rarer in cases of interpersonal crime such as serial homicide and rape, although some certainly exist. For example, most serial homicide cases rarely exceed 10 victims according to Hickey (1991) and we see similar lengths in cases of serial rape (e.g., 5.6 crimes on average in a sample of rapes collected by Canter and Larkin, 1993).

using risk assessment scores for sexual recidivism of sex offenders. To assess the extent to which these three statistics are appropriate (i.e., valid) measures of accuracy in this context, Hanson examined how the measures varied when two features of a sample were manipulated: the range of scores associated with the predictor variable(s) (actuarial risk scores in Hanson's case) and the base rates of the positive diagnostic alternative (recidivism in Hanson's case).

First, Hanson (2008) tested how the three measures of accuracy varied as a function of restriction of range (in the predictor variable) by dividing a sample of sex offenders who were assessed using the Static-99 risk assessment instrument (Harris, Phenix, Hanson, & Thornton, 2003) into those with low or high scores. By analyzing only low *or* high scoring offenders, Hanson could decrease the amount of variability in the predictor variable. By analysing low *and* high scoring offenders, Hanson could increase the amount of variability in the predictor variable. Results showed that manipulating range variability affected the measures of predictive accuracy when relying on *r* and the *AUC*, but the *OR* was not affected by this manipulation. Specifically, across the manipulations, *r* had a range of .37 (.12 to .49) and the *AUC* had a range of .18 (.60 to .78). The *OR* on the other hand remained relatively stable with a range of .12 (1.40 to 1.52). If these sorts of results generalize to the BLA context, they suggest that the *AUC* may be affected by variability in the predictor variable (across-crime similarity scores), which could be problematic.

The second manipulation Hanson (2008) examined involved changes to the base rates (incidents of recidivism). Hanson manipulated the base rates of recidivism by varying the length of follow up in years for his sample of offenders (i.e., a period of 1

year results in far less recidivism than a period of 16+ years). His results showed that while *r* decreases as the base rate decreases (from .33 for a 16+ years follow up to .21 for a 1 year follow up), there was no large change in either the *AUC* (.71 for 16+ years to .74 for a 1 year follow up) or the *OR* (1.49 for a 16+ years follow up to 1.52 for a 1 year follow up). This result for the *AUC* is not surprising given the discussion above regarding the *AUC* being unaffected by variability in the base rates (and previous research that has demonstrated the stability of this measure across base rate manipulations; e.g., Rice & Harris, 1995). However, the result for the *OR* was somewhat surprising given that the *OR* has been found to be somewhat unstable across base rate manipulations in previous research (e.g., Rice & Harris, 1995). However, at the very least, Hanson's study suggests that there may be value in at least considering the *OR* as a possible metric for assessing the predictive accuracy of tools used to make decisions in BLA given its potential robustness.

The Use of Similarity Coefficients in BLA

One of the reasons why it is important to have a valid measure of linking accuracy is so that researchers can appropriately assess the results of their studies, especially studies that examine factors that might impact the performance of decision making tools. One such factor that has started to receive attention from researchers is the choice of the similarity coefficient that is used to assess behavioural similarity and distinctiveness across a set of crimes (Bennell et al., 2010; Ellingwood et al., in press; Melnyk et al., 2010; Woodhams, Grant, et al., 2007).

In the majority of approaches to BLA, the similarity coefficient plays a central role. The general process for assessing BLA when using a similarity coefficient is fairly

straight forward. Two sub-samples are typically created from the full set of crimes being examined: one sample consists of all linked crime pairs and the other, all unlinked crime pairs. Then the chosen similarity coefficient is used to empirically assess the degree of behavioural similarity between each and every crime pair. High levels of across-crime similarity for crimes committed by the same offender represent greater degrees of behavioural stability, whereas low across-crime similarity scores committed by different offenders represent greater degrees of behavioural distinctiveness. When this pattern of results emerges (i.e., high similarity scores across crimes committed by the same offender and low similarity scores across crimes committed by different offenders), a high degree of linking accuracy can usually result (i.e., linked and unlinked crimes can be differentiated).

The problem is that many different similarity coefficients are available for use in BLA and it is not obvious which one is best. Compounding this problem is the fact that previous research in non-forensic areas such as biology and chemistry have demonstrated that the choice of similarity coefficient matters, in that it will influence accuracy in discriminatory tasks (Dalirsefat, Meyer, & Mirhoseini, 2009; Gower & Legendre, 1986; Kosman & Leonard, 2005; Sesli & Yegenoglu, 2010). For example, in a study examining the accuracy with which three similarity coefficients (Jaccard's, Sorensen-Dice, and simple matching) could distinguish between polymorphism markers in the silkworm, *Bombyx mori*, Dalirsefat, Meyer and Mirhoseini (2009) found that the similarity coefficients produced different results. This was also the case in a study by Sesli and Yegenoglu (2010) who examined the ability of the above three coefficients, along with

two others (Rogers and Tanimoto, and Russel and Rao), to distinguish between molecular markers of wild olives.

In the studies cited above, one of the most common coefficients examined was Jaccard's coefficient (Jaccard, 1908). Interestingly, in the forensic setting, the typical choice of similarity coefficient for the purposes of BLA is also Jaccard's coefficient (*J*). There are several reasons why *J* is the coefficient of choice for BLA. First, the appeal in part is due to the fact that *J* is easy to understand and easy to calculate. The calculation of *J* for a given crime pair (A and B) is simply:

$$J = \frac{a}{a+b+c}$$

where *a* equals joint occurrences of behaviour (1,1) across crimes A and B, *b* equals the number of times a behaviour is present in Crime A but not in Crime B (1,0), and *c* equals the number of times a behaviour is present in Crime B but not in Crime A (0,1). The value of *J* can range from 0 to 1, with 0 indicating perfect dissimilarity and 1 indicating perfect similarity.

Second, *J* is commonly used for the purposes of BLA because it excludes joint non-occurrences of behaviours, which would be represented by *d* (0,0 across Crimes A and B) (Bennell & Canter, 2002).[3] The assumption that is being made here by researchers is that joint non-occurrences of behaviour are not important for establishing the 'behavioural fingerprint' of an offender across his or her crimes because the absence of

[3] This is not to say that methods cannot be devised to include joint-absences of behaviour (i.e., joint non-occurrences) in the calculation of *J*. For example, it is possible to code variables in such a way that their presence in a dataset actually reflects the absence of a behaviour at a crime scene (e.g., where a positive coding of 1 indicates that cash was not stolen from a property despite it being available to steal). However, not only does such a procedure have the potential to cause confusion, it also makes assumptions that cannot be tested (e.g., that the offender actually did see the cash at the crime scene) and requires decisions to be made on the part of the data coder about which non-occurrences should be coded (a wide variety of behaviours are *not* exhibited by offenders when committing crimes, all of which could potentially be coded to reflect their absence).

behaviours could be due to reasons other than it not being exhibited by the offender (e.g., an error could be made by the police officer recording the crime, witness testimony could be unreliable or totally absent, victims may forget that the offender exhibited a behaviour or may be uncomfortable talking about certain behaviours with the police, etc.; Alison, Snook, & Stein, 2001; Woodhams & Labuschagne, 2011).

Third, a number of studies have now shown that Jaccard's coefficient can be used to discriminate between linked and unlinked crimes for a variety of crime types, including serial burglary (Bennell & Canter, 2002; Bennell & Jones, 2005; Melnyk et al., 2011 Markson et al., 2010), serial robbery (Woodhams & Toye, 2007), serial homicide (Melnyk et al., 2011; Santtila et al., 2008), serial arson (Ellingwood et al., in press), and serial sexual assault (Bennell et al., 2009; Bennell, Gauthier, et al., 2010; Woodhams, Grant, et al., 2007). For example, in a recent study by Woodhams and Labuschagne (2011), 119 rapes by 22 offenders from South Africa were examined using *J*. Their findings indicate that the use of *J* for BLA in cases of serial sex offenders can result in high levels of linking accuracy ($AUC = .88$).

Although Jaccard's coefficient has been the main similarity coefficient used to ascertain across-crime similarity in BLA, it is not the only similarity coefficient available for use. In fact, there are many similarity coefficients that are suitable for use with binary data, such as that used in the context of BLA (Dalirsefat, Meyer, & Mirhoseini, 2009; Liebetrau, 1983). However, it has only been recently that researchers have begun to question whether *J* is the best measure to use, and to compare the performance of *J* with other similarity measures to determine empirically whether one measure is better than another.

For example, Woodhams, Grant, et al. (2007) in a study of behavioural linkage of juvenile serial sex offenders, compared *J* to the taxonomic similarity index (Δ_S). They argued that *J* was too sensitive to slight variations in across-crime similarity because it accounts for similarity at only the most discrete levels. In contrast, Δ_S does not have that same sensitivity and therefore it should be a more robust coefficient for use in BLA. They argued that this is the case because Δ_S is not limited to the simple presence of crime scene behaviours across a pair of crimes. Instead, Δ_S organizes crime scene behaviours into a hierarchy such that, if across-crime similarity does not exist at the lowest (most discrete) level of behaviour (e.g., punching a victim) it can still be found to exist at higher (more general) levels of behaviour (e.g., exhibiting aggression; Bennell et al., 2010).

The findings of Woodhams, Grant, et al. (2007) provided some support for their hypothesis, but subsequent studies that have tried to replicate their work have not been successful. For example, a study of serial sex offenders by Bennell et al. (2010) did not find that Δ_S outperformed *J* in the ability to discriminate across crimes. The same was true in the study by Melnyk et al. (2011) where serial burglary and serial homicide were examined. The major problem with Δ_S in these studies appeared to be that it not only resulted in high similarity scores for crimes committed by the same offender, it also resulted in high levels of across-crime similarity for crimes committed by different offenders, thus making it potentially unsuitable for BLA purposes (Bennell et al., 2010; Melnyk et al., 2011).

The taxonomic similarity index is not the only similarity coefficient that has recently been compared to *J*, however. Recently, Ellingwood et al. (in press) examined how *J* compared with the simple matching coefficient (*S*) in an examination of serial

arson. This study was interesting because it gets to the heart of the key assumption for *J*'s use in BLA. Recall that one of the main reasons that researchers use *J* is that crime scene behaviours that are absent across two crimes (non-occurring) are deemed unimportant in establishing the offender's behavioural "signature" or behavioural "fingerprint" if you will. If this assumption is true, then the use of a similarity coefficient that excludes joint non-occurrences, such as *J*, is the appropriate choice. However, what if the opposite is correct? What if joint non-occurrences in behaviour across crimes are important? Presumably, the behaviours an offender does not commit may be as much an integral part of their behavioural fingerprint as the behaviours they do commit. If this is so, then by ignoring joint non-occurrences, we are ignoring important information about the offender and the crimes they commit that could increase behavioural linking accuracy.

The simple matching coefficient (*S*) is a coefficient that takes into account joint non-occurrences in behaviour. In fact, *S* has an equation very similar to that of *J* except that it includes *d*; occasions where joint non-occurrences of behaviours exist across Crime A and Crime B (0,0). This can be seen in the equation for *S* (Sokal & Michener, 1958):

$$S = \frac{a+d}{a+b+c+d}$$

The comparison of *J* and *S* by Ellingwood et al. (in press) found that *S* performed better than *J* at accurately discriminating between arsons committed by different offenders, but only once to a significant degree. That is to say that when examining both coefficients across three behavioural themes for arson (i.e., each of the three themes denoted a different offending style of arson), *S* performed as well as *J* on two of the themes, and outperformed *J* significantly in the case of the instrumental person arson

theme (the instrumental person theme is defined by the arson being linked to an emotional trigger usually involved with the breakdown of a relationship; Canter & Fritzon, 1998).

These recent findings raise the possibility that *S* might be a more appropriate measure than *J*, or at least as appropriate, but questions remain as to whether the results reported by Ellingwood et al. (in press) were due to the crime type being examined (serial arson). Thus, a replication of this study would be useful to see if *S* continues to perform as well as *J* (or better). If *S* were to significantly outperform *J*, it could hold important implications about the behavioural assumptions made in the choice of similarity coefficients for BLA. Specifically, it would suggest that the absence(s) of behaviour in crimes is an important part of understanding criminal behaviour, as it may suggest that what an offender does not do may be a crucial part of their crime signature or offending style.

There are also other commonly used similarity coefficients that might be useful for the purpose of BLA, such as the Sorensen-Dice (*SD*) index. This index is calculated in the following way (Dice, 1945; Sorensen, 1948):

$$SD = \frac{2a}{2a + b + c}$$

The *SD* index is particularly appealing in the current context because it also speaks directly to the issues just raised. Like *J*, this index ignores joint non-occurrences of behaviour across crimes and thus puts the focus more on occurrences of behaviour (though, like *J*, non-occurrences are still taken into account with *b* and *c*). However, in the case of *SD*, the focus on occurrences is even greater given that *SD* doubles the weight of joint occurrences.

In addition to establishing which accuracy measure should be used to assess performance in BLA, another goal of this study will be to use the accuracy measure determined to be most appropriate (in Phase 1 of this study) to compare the performance of *J* with other similarity coefficients, one which includes joint non-occurrences (*S*) and one, that like *J*, excludes joint non-occurrences of behaviour, but focuses even more on joint occurrences (*SD*). By comparing the performance between these three similarity coefficients (*J* vs. *S* vs. *SD*) across multiple crime types (serial burglary and serial rape in the present case), it will help us to better understand the conditions under which performance on linking tasks can be maximized. Determining which similarity coefficient performs best will also add to our understanding about the assumptions regarding behaviours within the context of BLA.

The Current Study

The main goals of the current study therefore are to expand on existing research on BLA by examining two primary issues. First, a comparative examination will be conducted of which accuracy measure is most appropriate for assessing performance in BLA: the correlation coefficient (*r*), the area under the curve (*AUC*), or the odds ratio (*OR*). The current study will essentially replicate the study by Hanson (2008) by comparing the three accuracy measures across the two manipulations he examined (restriction of range in the predictor variables and base rate) using two different crime types (serial burglary and serial rape). As discussed, this examination will help in determining what the most suitable (robust) accuracy measure is within the context of BLA.

Second, using the measure identified in the first phase of the study, this thesis will examine which similarity coefficient (*J*, *S*, or *SD*) leads to the highest levels of accuracy in cases of serial burglary and serial rape. This comparison will help add to our knowledge of which similarity coefficient is the most appropriate for use in BLA and will shed light on the effect of excluding joint non-occurrences of behaviour when conducting BLA, or relying more heavily on joint occurrences. Additionally, by using two different types of serial offences (property vs. interpersonal) to examine these issues, we can gain an understanding of how the choice of similarity coefficient can impact linking accuracy for different crime types. This could have an impact on our theoretical understanding of offender behaviour and on investigative practices regarding BLA.

Hypotheses

Hypothesis 1

The goal of Phase 1 in the current study will be to replicate Hanson's (2008) research, but within the context of BLA. It is expected that results of the current study comparing the sensitivity of *r*, the *AUC*, and the *OR* to variability in the predictor variable and variability in base rates will show similar results to that of Hanson's earlier study of risk decisions. That is, *r* and the *AUC* will prove more sensitive to variability in the predictor variable than the *OR*, while both the *AUC* and the *OR* will be less sensitive to variability in base rates compared to *r* (though Rice and Harris's (1995) study does open up the possibility that the *OR* will be impacted by the base rate manipulation).

Hypothesis 2

Phase 2 is more exploratory in nature, examining which of the similarity coefficients, *J*, *S*, or *SD*, results in better overall accuracy when conducting BLA (using

the measure identified as superior in Phase 1). There is some very preliminary research, which suggests that *S* may slightly outperform *J* (e.g., Ellingwood et al., in press), but that research is based on only a small sample of serial arsons. By exploring the results of this analysis across two different types of crime (serial burglary and serial rape) this study will allow for a more thorough examination of behavioural assumptions made when using BLA.

Methods

Samples

Serial burglary data. The serial burglary data was originally collected by crime analysts and entered directly into a data management system. The serial burglary data used in the current study represents a subset of the above data, which was originally collected by Bennell (2002) for research purposes. The data contains 28 crime scene behaviours from 210 solved residential burglaries committed in the UK by 42 male serial burglars. Each behaviour was coded as present (1) or absent (0) in each crime. As the data was archival it was not possible to assess inter-rater reliability. However, previous research has indicated that this type of crime scene data can be coded reliably (Alison & Stein, 2001; Häkkänen, Puolakka, & Santtila, 2004). The data was limited to five crimes per offender. This restriction is common practice in BLA research in order to ensure that prolific offenders (who may be very stable or unstable) do not bias the results (Bennell et al., 2010; Bennell & Jones, 2005; Woodhams & Toye, 2007). See Appendix A for a list of all the behaviours included in this data set.

Serial rape data. The serial rape[4] data used in the current study represents a subset of data originally collected for previous serial sex offender research (Canter, Wilson, Jack, & Butterworth, 1996). The data contains information on 36 crime scene behaviours from 126 solved serious sexual assault offences committed by 42 adult male offenders in the UK. All victims were female and did not know their attacker. The serial rape data was originally collected through the records of victim statements from various police forces across the UK. The behaviours included in the victim statements were coded by previous researchers. Each behaviour was coded as present (1) or absent (0) for each crime. It was not possible to assess inter-rater reliability for reasons discussed above. The data was restricted to three crimes per offender for the purpose of the original study (Canter et al., 1996). As indicated above, this is normal practice in BLA research. See Appendix B for a list of all the behaviours included in this data set.

Procedure

The current study is comprised of two phases. Each phase has a number of steps, which are described in more detail below. Each step will be repeated twice: once for the serial burglary data and a second time for the serial rape data.

Phase 1: Procedure for comparing *r*, the *AUC*, and the *OR*.

Manipulation of variability in the predictor variable. Each sample of crimes (burglary and rape) consists of a data matrix of rows, which signify crimes (e.g., crime 1-1, which would indicate the first offence committed by offender 1) and columns, which signify behaviours (e.g., entered property through the front door). The cells of the data matrix consist of 1's or 0's, which signify whether a particular behaviour was present or

[4] Note that crimes were defined by using British crime definitions of rape and that these definitions of rape can change depending on jurisdiction/state/country.

absent in a particular crime. The data matrix will be submitted to a specially designed computer program referred to as B-LINK (Bennell, 2002). This program takes as input the data matrix just described and produces as output a listing of each crime pair in the sample, an indication for each crime pair as to whether the crimes were committed by the same offender (linked) or different offenders (unlinked), and similarity scores (Jaccard's coefficient, simple matching, and Sorensen-Dice), which range from 0 to 1 for each pair of crimes (indicating the degree of behavioural similarity that exists across a crime pair). Prior to the data being split for the purpose of examining hypothesis 1, similarity scores for linked and unlinked crimes will be examined to assess if linked crimes show significantly greater similarity than unlinked crimes.

The output file from B-LINK will then be split into quartiles based on the similarity coefficients to reflect crime pairs characterized by very low across-crime similarity, moderately low levels of across-crime similarity, moderately high levels of across-crime similarity, and high levels of across-crime similarity. These quartiles will form new sub-samples of crime pairs that will allow us to test the impact of range restriction on the various accuracy measures under consideration (*r*, the *AUC*, and the *OR*). Specifically, in line with Hanson (2008), we can *decrease* the variability of scores by selecting, for example, only crime pairs with very low scores *or* high scores. Alternatively, we can *increase* the variability of scores by combining extreme groups by selecting, for example, crime pairs with very low scores *and* high scores. By examining the results across various combinations of samples, it can be determined how each of the three accuracy measures varies as a function of range restriction (samples will be combined specifically to increase the degree of variability). Also, the correlation between

the standard deviations of the samples used (indicating the degree of range restriction, or variability, for the sample) and the values of the accuracy measures will be calculated, as per Hanson (2008).

The accuracy measure, *r*, will be calculated for each sub-sample by correlating each of the similarity coefficients with the linked/unlinked variable (obtained from the original B-LINK output file). The measure, *r*, in this case is the point-biserial correlation between these variables. The *AUC* will be calculated by constructing a ROC graph using the similarity coefficients as the test variables and the linked/unlinked variable as the state variable. The *AUC* in this case refers to the probability that a randomly selected linked crime pair has a higher similarity score than a randomly selected unlinked crime pair. The *OR* will be calculated by constructing logistic regression models for each of the samples and extracting the *OR*. The *OR* tells you the proportionate change in odds of a crime pair being linked given a unit change in the predictor variable (i.e., the degree of across-crime similarity in the current case) (Field, 2009). An *OR* greater than 1 indicates that, with an increase in the predictor variable, the odds of the crime pair being linked increases (the reverse is true for *OR*s < 1). SPSS v. 20 will be used to calculate all accuracy measures.

Manipulation of variability in the base rates. Using the output file from B-LINK (described above), sub-samples consisting of different base rates of randomly sampled linked and unlinked crimes will also be constructed. Each sample will include a pre-determined percentage of linked and unlinked crime pairs, an indication of which crimes are linked and which are unlinked, and the similarity coefficient (*J*, *S*, and *SD*) associated with each pair indicating the degree of similarity exhibited across the crimes.

Specifically, five different base rate combinations will be examined: (1) 50% linked crimes/50% unlinked crimes, (2) 40% linked crimes/60% unlinked crimes, (3) 30% linked crimes/70% unlinked crimes, (4) 20% linked crimes/80% unlinked crimes, and (5) 10% linked crimes/90% unlinked crimes. To increase the reliability of the results, 10 random samples representing each base rate combination from (1) to (5) will be constructed with average results (for each of the accuracy measures under consideration) being calculated across the samples. To combine the results for each set of 10 random samples, the standard errors (SE) and 84% confidence intervals (CI_{84})[5] will be combined using the procedure outlined by Borenstein, Hedges, and Rothstein (2007).[6] For each sub-sample of crime pairs, the three measures of predictive accuracy will be calculated and interpreted as described above.

Phase 2: Procedure for comparing *J*, *S*, and *SD*.

Using B-LINK to calculate J, S, and SD. In order to compare the ability of *J*, *S*, and *SD* to link serial burglars and rapists, the following steps will be taken. The data matrices for serial burglary and rape, which were described above, will be subjected to B-LINK. As above, the output file from B-LINK will create crime pairs, indicate which pairs are linked or unlinked, and provide measures of across-crime similarity for each

[5] When the overlap of CIs are going to be used to determine if differences between scores are significantly different, the use of 84% CIs is often suggested to approximate an alpha level of .05 (Payton, Greenstone, & Schenker, 2003; Tryon, 2001).The use of traditional 95% CIs are considered by some to be too conservative for detecting type 1 errors.

[6] It is important to note that the 10 random samples that are being averaged are non-independent (i.e., some of the same crimes and same offenders will appear across some of the 10 random samples). Although it would be ideal to use 10 independent samples for this purpose, this was not possible due to the size of the original sample (i.e., only 42 offenders per crime type). The result of drawing on samples that are not independent is that the CIs that emerge from these samples are likely to be narrower (Bonett, 2009). This will result in a greater likelihood of concluding that differences in accuracy measures exist across the manipulations.

pair. For the purposes of the present study, *J, S,* and *SD* will be calculated using all 28 behaviours for serial burglars and all 36 rape behaviours.

For the purposes of our analysis, two separate output files will be used. The first file (termed the unequal file) is the one that emerges from B-LINK. This output file will have an unequal number of linked and unlinked crime pairs (given that B-LINK creates all possible crime pairings from a sample of serial crimes, there will always be many more unlinked crime pairs compared to linked crime pairs in the output file). Similar types of files have been used in previous linking research (e.g., Bennell & Canter, 2002; Bennell & Jones; 2005; Bennell et al., 2010; Ellingwood et al., in press; Melnyk et al., 2011) and they represent reasonably well what the police would be faced with when conducting BLA in naturalistic settings. The second file (termed the equal file) will be manipulated so that the number of linked and unlinked crime pairs is equal (the unlinked crimes will be randomly chosen to make the number of unlinked crime pairs equal to the number of linked crime pairs). This procedure has also been used in previous linking research (e.g., Markson et al., 2010; Tonkin et al., 2008; Woodhams, Grant, et al., 2007; Woodhams & Toye, 2007) based on the assumption that it may be important to reduce the potentially biasing effect associated with having too many unlinked crime pairs. The current study represents the first attempt to empirically assess the impact of these two sampling procedures.

Split-half validation. Prior to examining the differences in predictive accuracy that emerge when using these three similarity coefficients, the equal and unequal files described above will be split in order to perform split-half validation. Split-half validation will be used to reduce potential biases that can occur when the same data is used to

develop and test prediction models (Efron, 1982). In the current case, both the equal and unequal samples will be split by randomly selecting 50% of the linked crime pairs in each sample and 50% of the unlinked crime pairs. The samples that will be used to develop the prediction models will be referred to as the development samples. The samples used to test the prediction models will be referred to as the validation samples.

Evaluating regression models based on J, S, and SD. Logistic regression models will be constructed for *J, S,* and *SD* using data in the developmental samples (for both the equal and unequal files). Logistic regression is an appropriate analytical technique to use in this case given the dichotomous nature of the criterion variable (linked/unlinked; Tabachnick & Fidell, 2007). These models will then be used to calculate predicted probabilities of crime pairs being linked in the test samples (for both the equal and unequal files). The accuracy of these regression models will then be assessed using the measure of accuracy (*r*, the *AUC*, or the *OR*) that proved to be the most robust in Phase 1 of the current study. Differences between the similarity coefficients will be assessed by comparing the 84% CIs associated with the measure of accuracy.

Results

Phase 1

Descriptive and comparative analyses. Descriptive statistics were calculated for each of the similarity coefficients (*J, S,* and *SD*) across all linked and unlinked crime pairs for the rape and burglary data sets (Table 1). To ensure that the similarity scores for the linked crimes differed significantly from unlinked crimes, significance tests of the differences between the distributions of linked and unlinked crime pairs were conducted for each similarity measure. As expected given the results of previous BLA research,

tests of normality indicated that the distributions of the similarity scores for linked and unlinked distributions were not normal for each of the similarity coefficients (all p's < .001). Therefore non-parametric tests were used to conduct the tests of significance between the distributions.

As can be seen in Table 2, the comparison of linked versus unlinked distributions show that the means are higher for linked distributions than unlinked distributions regardless of which similarity coefficient is used. These results demonstrate that the levels of behavioural stability observed across crimes committed by the same offender are higher than the levels of behavioural stability observed across crimes committed by different offenders. This is consistent with previous BLA research (e.g., Bennell & Jones, 2005; Bennell et al., 2009; Goodwill & Alison, 2006; Woodhams, Grant, et al., 2007; Woodhams & Labuschagne, 2011). The results in Table 2 also show that the highest scores, for both linked and unlinked crime pairs, are associated with *S*. The scores associated with *SD* are slightly lower than those associated with *S*, and the lowest scores are associated with *J*. Thus, across-crime similarity scores do seem to vary as a function of the similarity coefficient used.

Table 1

Descriptive Statistics and Significance Tests for the Linked and Unlinked Distributions of J, S, and SD Scores for Rape and Burglary Samples

Coefficient		Range		Median		Mean (*SD*)		Wilcoxon	*p*-value	Effect size
		L	UL	L	UL	L	UL			
J	Rape	.08-.80	.00-.86	.39	.22	.39 (.16)	.23 (.11)	-8.70	<.001	.77
	Burglary	.00-1.00	.00-1.00	.27	.19	.29 (.19)	.22 (.17)	-4.72	<.001	.23
S	Rape	.58-.97	.36-.97	.81	.69	.98 (.08)	.69 (.09)	-7.44	<.001	.66
	Burglary	.46-1.00	.32-1.00	.75	.71	.75 (.10)	.71 (.10)	-1.70	<.100	.08
SD	Rape	.14-.89	.00-.92	.56	.36	.54 (.17)	.36 (.15)	-8.61	<.001	.77
	Burglary	.00-1.00	.00-1.00	.43	.32	.42 (.22)	.33 (.32)	-4.57	<.001	.22

Note. L: linked crime pairs for rape (n = 126); UL: unlinked crime pairs for rape (n = 7749); L: linked crime pairs for burglary (n = 420); UL: unlinked crime pairs for burglary (n = 21525); SD: standard deviation; effect size = $r = z/\sqrt{N}$ (.00–.30 = small effect; .30–.50 = moderate effect; .50– = large effect).

Table 1 also presents the Wilcoxon tests of significance, which confirm that significant differences exist between the linked and unlinked distributions of similarity scores for all comparisons with the exception of *S* for burglary. Interestingly, the effect sizes associated with these differences, which are also included in Table 1, indicate that the differences in across-crime similarity scores between linked and unlinked crimes are much smaller for burglary than for rape. These results suggest that BLA may be more feasible in cases of rape compared to burglary, although the overlap in the distributions revealed by the respective ranges of similarity scores suggest that perfect linking accuracy will not be achieved with either data set.

To illustrate further how the distributions of similarity scores for linked and unlinked crimes overlap, histograms were constructed (see Figures 1 through 6). Since the across-crime similarity scores across these distributions do overlap, it indicates that perfect linking accuracy will not be possible regardless of which similarity score is used or which data set is examined (rape or burglary). However, consistent with the results in Table 1, the figures do illustrate that the linked and unlinked distributions of similarity scores for the rape data overlap less than they do for the burglary data, which suggests that linking accuracy scores will be higher for cases of rape in the current study.

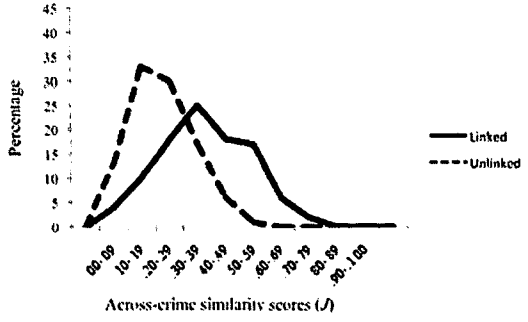

Figure 1. Distributions of across-crime similarity scores for linked and unlinked rapes using *J*.

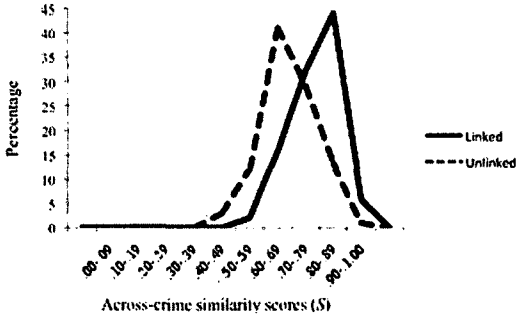

Figure 2. Distributions of across-crime similarity scores for linked and unlinked rapes using *S*.

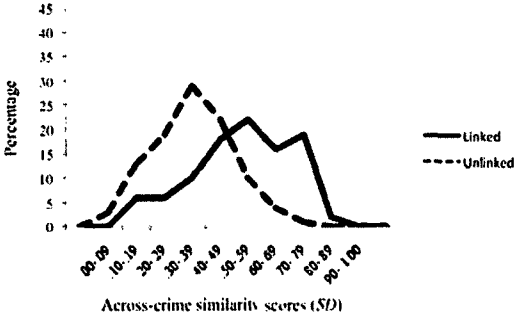

Figure 3. Distributions of across-crime similarity scores for linked and unlinked rapes using *SD*.

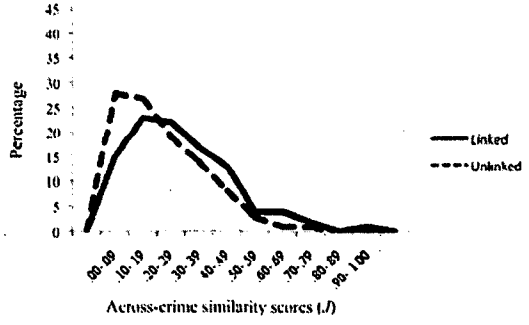

Figure 4. Distributions of across-crime similarity scores for linked and unlinked burglaries using *J*.

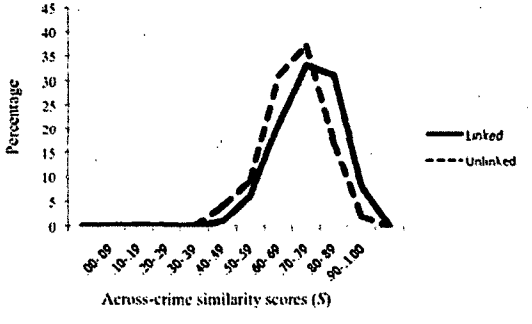

Figure 5. Distributions of across-crime similarity scores for linked and unlinked burglaries using *S*.

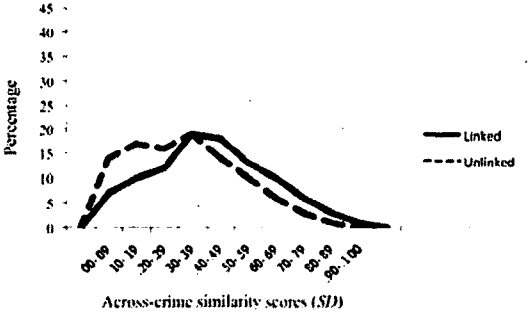

Figure 6. Distributions of across-crime similarity scores for linked and unlinked burglaries using *SD*.

The impact of restriction of range on *r*, the *AUC*, and the *OR*. To determine which accuracy measure(s) (*r*, the *AUC*, or the *OR*) is/are most robust, analyses were conducted to examine the sensitivity of the measures to variability in the predictor variable (similarity scores). To do this, restriction of range in the similarity scores was artificially increased or decreased using the procedure employed by Hanson (2008), and the impact of this change on the accuracy measures was observed.

In examining Table 2, which presents the results for serial rape, it can be seen that variability in the similarity scores (*J*, *S*, and *SD*) resulted in changes to the accuracy measures, regardless of which accuracy measure was used, though significant changes were observed more frequently for *r* and the *AUC* than for the *OR*. Indeed, for these two measures, there are numerous instances where non-overlapping CIs are observed, indicating that significant differences exist between the accuracy scores obtained for particular sub-samples of rape data as a function of variability in the predictor variable. These results were expected based on Hanson's (2008) study.

Table 2

Changes in the Magnitude of Prediction Statistics Based on Restriction of Range of the Predictor Variable (Similarity Coefficient Scores) for Rape Data Sub-samples

Scores	Standard Deviation			r (CI_{84})			AUC (CI_{84})			OR (CI_{84})		
	J	*S*	*SD*	*J*	*S*	*SD*	*J*	*S*	*SD*	*J*	*S*	*SD*
Low[a]	.06	-	.07	.02 (-.00-.04)	-	.04 (.01-.07)	.56 (.48-.64)	-	.63 (.54-.72)	4.20×10^{1} (4.7×10^{-1}-3.78×10^{3})	-	1.16×10^{4} (2.74×10^{0}-4.95×10^{7})
Mod Low[b]	.06	.03	.07	.09 (.06-.12)	-	.05 (.03-.07)	.66 (.61-.71)	-	.66 (.61-.72)	3.59×10^{3} (3.02×10^{2}-4.27×10^{4})	-	4.66×10^{3} (1.45×10^{2}-1.49×10^{5})
Mod High	.05	.06	.06	.15 (-.03-.27)	.05 (.03-.07)	.19 (.16-.23)	.63 (.55-.70)	.67 (.61-.72)	.68 (.63-.74)	7.73×10^{2} (3.51×10^{0}-1.70×10^{5})	7.44×10^{4} (1.18×10^{3}-4.68×10^{6})	1.96×10^{5} (1.23×10^{4}-3.15×10^{6})
High[c]	.04	.04	.04	-	.11 (.08-.15)	.05 (-.20-.29)	-	.64 (.60-.69)	.48 (.30-.65)	-	7.98×10^{4} (2.67×10^{3}-2.39×10^{6})	1.44×10^{1} (.00-1.34×10^{7})
All	.12	.09	.15	.17 (.16-.19)	.13 (.12-.15)	.15 (.14-.17)	.78 (.75-.82)	.77 (.74-.80)	.78 (.75-.82)	7.47×10^{3} (3.02×10^{3}-1.85×10^{4})	1.38×10^{5} (2.97×10^{4}-6.42×10^{5})	3.72×10^{3} (1.50×10^{3}-9.26×10^{3})
Low & Mod Low[d]	.11	.03	.12	.09 (.07-.11)	-	.04 (.02-.06)	.72 (.68-.76)	-	.63 (.57-.68)	2.05×10^{3} (5.35×10^{2}-7.92×10^{3})	-	5.94×10^{1} (8.56×10^{0}-4.12×10^{2})
Low, Mod Low &	.12	.06	.15	.17 (.16-.19)	.05 (.03-.07)	.13 (.11-.15)	.78 (.75-.82)	.68 (.63-.73)	.77 (.73-.80)	8.75×10^{3} (3.45×10^{3}-2.22	1.08×10^{5} (1.92×10^{3}-6.06	2.88×10^{3} (1.04×10^{3}-8.02

Scores	Standard Deviation			r (CI_{84})			AUC (CI_{84})			OR (CI_{84})		
	J	*S*	*SD*	*J*	*S*	*SD*	*J*	*S*	*SD*	*J*	*S*	*SD*
Mod High										x 10^4)	x 10^6)	x 10^3)
Mod Low & High	.06	.12	.08	.10 (.07-.13)	.11 (.08-.14)	.15 (.13-.17)	.66 (.61-.71)	.69 (.65-.73)	.73 (.68-.78)	2.23×10^3 (2.78×10^2-1.79 x 10^4)	1.20×10^5 (5.08×10^3-2.87 x 10^6)	1.15×10^4 (2.85×10^3-4.71 x 10^4)
Low & High[e]	.07	.04	.11	.04 (.03-.05)	.11 (.08-.15)	.29 (.26-.33)	.58 (.46-.69)	.64 (.60-.69)	.79 (.72-.86)	4.13×10^2 (2.97×10^1-5.75 x 10^3)	7.98×10^4 (2.67×10^3-2.39 x 10^6)	8.07×10^2 (2.74×10^2-2.38 x 10^3)
r[f]				.63	.27	.28	.86**	.64	.73*	.78*	.78*	-.34

Note. Similarity scores of rape data samples selected: Low (.00-.25); Moderate Low (.26-.50); Moderate High (.51-.75); High (.76-1.00); All (.00-1.00); Low and Moderate Low (.00-.50); Low, Moderate Low, and Moderate High (.00-.75), Moderate Low and High (.26-.50 and .76-1.00), Low and High (.00-.25 and .76-1.00); *r*, correlation coefficient; *AUC*, area under the curve; *OR*, odds ratio; CI_{84}, 84% confidence intervals.

[a]No samples scored ≤.25 for *S*. [b]No moderate low samples were linked for *S*. [c]Could not compute for *J* due to small sample size ($n = 3$) in which there was only one linked pair ($n = 1$). [d]No low or moderate low linked samples available for *S*. [e]Calculated without low range for *S* as no samples scored ≤.25. [f]Correlations between standard deviations and accuracy levels.

*$p < .05$. **$p < .01$.

While it is tempting to place importance on the *OR* results found here, the values of the *OR*s that were obtained in this analysis, while correct, are very difficult to interpret (as are the very wide CIs associated with the *OR*s). What can account for the very extreme values that were observed in this analysis? Arguably, the relatively unusual nature of the data used in BLA research is largely responsible for the observed *OR*s. Two issues seem particularly problematic (Craig Leth-Steensen[7], personal communication, July 14, 2012).

First, the base rate of linked crimes is extremely low (<2%) in all of the sub-samples that were used for these analyses. Second, and perhaps more importantly, the maximum range of the predictor variable in the current case is 1 unit, though in most of the rows included in Table 2 the range is even less than that.[8] What appears to explain the large values of the *OR* then is the fact that the odds of two crime pairs being linked is exceptionally small at low levels of the predictor variable (i.e., 0). Given that these small odds form the denominator of the *OR*, the potential for extremely large *OR*s becomes apparent (Craig Leth-Steenson, personal communication, August 27, 2012).

In addition to this issue, it should also be noted (as is done in the notes attached to Table 2) that several unexpected problems were encountered when attempting to calculate the values of the similarity scores for some of the sub-samples. Specifically,

---

[7] Dr. Craig Leth-Steenson, Statistics Consultant, Department of Psychology, Carleton University.

[8] The fact that the range for the similarity scores in most of the rows in Table 2 is <1 might even mean that the *OR*s for those rows are not able to be interpreted (Craig Leth-Steensen, personal communication, August 27, 2012). A minimum 1-unit increase in the predictor variable may be required in order for an *OR* to have meaning. If the maximum range for a predictor variable is 1, that predictor variable would presumably be treated as a categorical (0,1) variable when interpreting *OR*s (Craig Leth-Steenson, personal communication, August 27, 2012). This line of reasoning was confirmed when we scaled the predictor variables by multiplying their values by 100. The impact of this scaling procedure was that *r* and the *AUC* did not change for any sub-sample listed in Table 2, but the values for the *OR*s (and their CIs) were less extreme (i.e., in the "normal" range for *OR*s). While I considered running all analyses from this point forward with this scaled data, ultimately it was felt that the purpose behind the analyses in this thesis was to explore the data that is typically used by researchers in the context of BLA research, which are similarity scores that range from 0 to 1.

problems with *J* were encountered when trying to calculate *r* and the *AUC* for the high similarity score sub-sample because there were only 3 crime pairs in that sub-sample, of which only one pair was linked. Thus, unreliable results were obtained. Unlike the problems encountered for *J*, in the case of *S*, the problem was that no similarity scores existed for some of the sub-samples (specifically those that were meant to contain low similarity scores). This makes sense given how *S* is calculated; that is, the fact that *S* includes joint non-occurrences in its calculation would lead one to expect relatively high values compared to the other two measures (*J* and *SD*).

With respect to the correlations calculated for each of the accuracy measures (based on the accuracy scores and standard deviations associated with each sample) the results were not entirely as expected. For example, with respect to *r*, it was expected (based on Hanson, 2008) that values of this accuracy measure would positively correlate with variability in the predictor variable, such that higher values of *r* would be observed for samples with larger standard deviations. However, a high value of *r* was only found when using *J* ($r = .63$), and although it was a large and positive value, the correlation was not statistically significant. With respect to the *AUC*, significant positive correlations were also expected based on Hanson's study. In this case, expectations were confirmed for *J* ($r = .86$) and *SD* ($r = .73$), but the correlation for *S* ($r = .64$), while being large and positive, did not reach statistical significance. These findings indicate that the *AUC* therefore was significantly impacted by range restriction, while *r*, though affected (as indicated by non-overlapping CIs in a number of instances) was not affected to a statistically significant degree. Finally, with respect to the *OR*, non-significant correlations were expected based on past research, but the correlations for *J* ($r = .78$) and

S ($r = .78$) were both significant. Given the extreme values of the *OR*s (and their CIs) that were found in this study, and the probable explanations for these values, caution is warranted when interpreting these results.

Similar analyses were conducted for serial burglary. Odds ratios were calculated once again to ensure that the extreme results observed in the previous analyses were not unique to the serial rape dataset. As illustrated in Table 3, very similar results to the rape analyses were found in the case of serial burglary. More specifically, it can be seen that variability in the restriction of range (in *J*, *S*, and *SD*) for burglary resulted in changes to the accuracy measures, regardless of which accuracy measure was used, but it was only for *r* and the *AUC* that non-overlapping CIs were found. As mentioned above, these results are not unexpected given the findings of Hanson (2008). However, once again, the values of the *OR*s, while correct, are very extreme, making the interpretation of these results difficult (though it should be pointed out that the *OR*s in this table, and their CIs, are not as extreme as those that are reported in Table 2).

As happened in the analysis of the serial rape data, we encountered several unexpected problems when attempting to calculate the similarity scores for some of the sub-samples. Specifically, like with rape, no similarity scores for *S* existed for the low score sub-sample.

Table 3

Changes in the Magnitude of Prediction Statistics Based on Restriction of Range of the Predictor Variable (Similarity Coefficient Scores) for Burglary Data Sub-samples

Scores	Standard Deviation			r (CI_{84})			*AUC* (CI_{84})			*OR* (CI_{84})		
	J	*S*	*SD*	*J*	*S*	*SD*	*J*	*S*	*SD*	*J*	*S*	*SD*
Low[a]	.08	-	.10	.03 (.02-.05)	-	.01 (-.01-.03)	.58 (.55-.61)	-	.54 (.50-.58)	32.29 (9.07-114.93)	-	3.37 (.74-15.28)
Mod Low	.07	.03	.09	.02 (.00-.04)	.04 (-.01-.09)	.03 (.02-.05)	.53 (.49-.56)	.63 (.49-.78)	.54 (.51-.57)	4.49 (0.99-20.31)	1.81×10^{12} ($.00-3.72 \times 10^{28}$)	7.09 (1.71-29.48)
Mod High	.06	.06	.06	.06 (.01-.11)	.03 (.02-.04)	.01 (-.01-.03)	.60 (.53-.66)	.56 (.54-.59)	.52 (.48-.56)	117.30 ($2.94-4.67 \times 10^3$)	38.80 (7.88-191.07)	2.69 (.32-22.54)
High	.08	.04	.06	.12 (-.04-.27)	.05 (.03-.07)	.07 (-.00-.14)	.54 (.35-.72)	.57 (.54-.60)	.55 (.46-.63)	76.31 ($.23-2.58 \times 10^4$)	223.35 ($27.80-1.79 \times 10^3$)	89.59 ($1.27-6.34 \times 10^3$)
All	.16	.10	.21	.06 (.05-.07)	.06 (.05-.07)	.06 (.05-.07)	.62 (.60-.64)	.62 (.60-.64)	.62 (.60-.64)	11.75 (8.13-16.96)	76.03 (37.59-153.80)	8.03 (5.79-11.13)
Low & Mod Low[b]	.14	.03	.16	.05 (.04-.06)	.04 (-.01-.09)	.04 (.03-.05)	.61 (.59-.63)	.63 (.49-.78)	.60 (.57-.62)	13.87 (8.27-23.25)	1.81×10^{12} ($.00-3.72 \times 10^{28}$)	8.85 (4.96-15.78)
Low, Mod Low & Mod High[b]	.16	.08	.20	.06 (.05-.07)	.03 (.02-.04)	.05 (.04-.06)	.62 (.59-.64)	.57 (.55-.60)	.61 (.59-.63)	10.83 (7.28-16.13)	52.19 (12.80-212.88)	6.45 (4.47-9.31)

Scores	Standard Deviation			r (CI_{84})			*AUC* (CI_{84})			*OR* (CI_{84})		
	J	*S*	*SD*	*J*	*S*	*SD*	*J*	*S*	*SD*	*J*	*S*	*SD*
Mod Low & High	.09	.12	.12	.05 (.03-.07)	.06 (.04-.08)	.06 (.05-.08)	.54 (.51-.58)	.60 (.57-.63)	.58 (.55-.61)	11.22 (4.45-28.28)	145.27 (35.29-597.99)	12.68 (6.92-23.26)
Low & High[b]	.10	.04	.18	.06 (.05-.07)	.05 (.03-.07)	.09 (.07-.11)	.59 (.57-.62)	.57 (.54-.60)	.62 (.58-.66)	19.59 (10.64-36.05)	223.35 (27.80-1.79 x 10^3)	11.80 (7.65-18.18)
r[c]				.05	.46	.47	.71*	-.04	.94**	-.55	-.59	-.41

Note. Similarity scores of burglary data samples selected: Low (.00-.25); Moderate Low (.26-.50); Moderate High (.51-.75); High (.76-1.00); All (.00-1.00); Low and Moderate Low (.00-.50); Low, Moderate Low, and Moderate High (.00-.75), Moderate Low and High (.26-.50 and .76-1.00), Low and High (.00-.25 and .76-1.00); r, correlation coefficient; *AUC*, area under the curve; *OR*, odds ratio; CI_{84}, 84% confidence intervals.
[a]No samples scored ≤.25 for *S*. [b]Calculated without low range for *S* as no samples scored ≤.25. [c]Correlations between standard deviations and accuracy levels.
*$p < .05$. ** $p < .01$.

The results from the correlational analyses that are reported in Table 3 were not completely as expected. For example, with respect to *r*, significant positive correlations were expected. However, non-significant correlations were found across the three similarity coefficients (*J, S,* and *SD*). Regarding the *AUC*, significant positive correlations were again expected, but only the correlations associated with *J* ($r = .71$) and *SD* ($r = .94$) were found to be significantly positive. These findings suggest that the *AUC* was more significantly affected by restriction of range in the predictor variable than *r*. Finally, based on previous research, non-significant correlations were expected for the *OR* and this was found to be the case. However, for all the reasons discussed above, caution is warranted when interpreting the results for the *OR*.

Effect of base rate variability on *r* and the *AUC*. The initial plan for the examination of base rate variability was to investigate all three accuracy measures (*r*, the *AUC*, and the *OR*). However, given the extreme values of the *OR*s in the previous analyses, and the very wide CIs associated with the *OR*s, a decision was made to exclude the *OR* from this analysis. The rest of this thesis will focus solely on *r* and/or the *AUC*.

To test the sensitivity of *r* and the *AUC* to variability in the base rate, base rates for serial rape and burglary were manipulated, as per Hanson (2008), by artificially decreasing the number of linked crimes included in a sub-sample while keeping the overall size of the sub-samples to be tested the same. Tables 4 and 5 indicate to what degree the chosen accuracy measures (*r* and the *AUC*) are affected by manipulations of the base rates for serial rape and burglary, respectively.

The results reported for rape in Table 4 demonstrate that, regardless of which similarity coefficient is examined, *r* decreases as the base rate of linked crimes decreases

(e.g., from .17 to .06 as the base rate is varied from 50% to 10% for *J*). This result is similar to the findings of Hanson (2008). While the CIs associated with *r* did overlap across many of the sub-samples, the CIs did not overlap between the 50/50 and the 10/90 sub-samples, suggesting that *r* is significantly affected by base rate variability, at least in some instances. However, the difference between the non-overlapping CIs in the above instance was narrow and thus should be interpreted with caution since the samples were non-independent (the difference was .01 between the two CIs; see footnote 5).

Table 4

Effect of Variability of Base Rates on the Magnitude of Effect Sizes (r and AUC) Using Rape Data

Base Rate Manipulation[a]	*r* (CI_{84})			*AUC* (CI_{84})		
	J	*S*	*SD*	*J*	*S*	*SD*
50/50	.17 (.12-22)	.13 (.08-.17)	.15 (.08-.22)	.78 (.73-.84)	.77 (.73-.81)	.78 (.72-.85)
40/60	.13 (.08-.19)	.10 (.05-.14)	.12 (.05-.19)	.78 (.72-.83)	.76 (.72-.80)	.77 (.70-.84)
30/70	.12 (.07-.17)	.09 (.04-.13)	.10 (.04-.17)	.79 (.74-.84)	.78 (.73-.82)	.79 (.72-.86)
20/80	.08 (.03-.13)	.06 (.02-.11)	.07 (.00-.14)	.75 (.70-.81)	.76 (.72-.80)	.75 (.69-.82)
10/90	.06 (.01-.11)	.05 (.00-.09)	.05 (-.02-.12)	.79 (.73-.84)	.80 (.75-.84)	.79 (.70-.85)

Note. [a]Linked/unlinked ratio; *r*, correlation coefficient; *AUC*, area under the curve; CI_{84}, 84% confidence intervals.

In contrast to *r*, the *AUC* values are really not affected at all by the base rate fluctuations, regardless of what similarity measure is used (e.g., from .78 to .79 as the base rate is varied from 50% to 10% for *J*). Not only do the CIs overlap across the various samples that were tested for each of the similarity coefficients, but the CIs remain nearly identical across these samples. Again, this result is consistent with the findings of Hanson (2008) and suggests that the *AUC* is a more robust measure for use in BLA with respect to changes in the base rate of linked crimes. One clearly does not want to use an accuracy measure that varies significantly as a function of base rate.

The results reported for serial burglary in Table 5 indicate that the correlation coefficient decreases as the base rate of linked crimes decreases (e.g., from .06 to .02 as the base rate is varied from 50% to 10% for *J*) regardless of which similarity coefficient is used. However, all of the CIs associated with the *r* values for serial burglary overlapped, suggesting that base rate fluctuation does not appear to affect *r* to a significant degree for the current sample of burglaries. This runs counter to what was expected based on Hanson's (2008) results.

Table 5

Effect of Variability of Base Rates on the Magnitude of Effect Sizes (r and AUC) Using Burglary Data

Base Rate Manipulation[a]	*r* (CI_{84})			*AUC* (CI_{84})		
	J	*S*	*SD*	*J*	*S*	*SD*
50/50	.06 (-.01-.13)	.06 (.01-.11)	.06 (-.04-.15)	.61 (.54-.69)	.62 (.57-.66)	.61 (.52-.71)
40/60	.05 (-.02-.13)	.05 (.00-.10)	.05 (-.05-.15)	.62 (.55-.69)	.62 (.57-.66)	.62 (.52-.72)
30/70	.04 (-.03-.12)	.04 (-.01-.09)	.04 (-.06-.14)	.62 (.55-.70)	.62 (.57-.66)	.62 (.53-.72)
20/80	.03 (-.04-.11)	.03 (-.02-.08)	.03 (-.07-.13)	.61 (.54-.69)	.62 (.58-.67)	.61 (.52-.71)
10/90	.02 (-.05-.09)	.02 (-.03-.07)	.02 (-.08-.11)	.60 (.52-.67)	.60 (.55-.69)	.60 (.50-.69)

Note. [a]Linked/unlinked ratio; *r*, correlation coefficient; *AUC*, area under the curve; CI_{84}, 84% confidence intervals.

Once again, the *AUC* values were not affected at all by the base rate fluctuations regardless of what similarity coefficient was used (e.g., from .61 to .60 as the base rate is varied from 50% to 10% for *J*), which is consistent with Hanson's (2008) conclusion. As demonstrated by the results in Table 5, the CIs overlapped across the various base rate manipulations for each of the similarity coefficients, suggesting that the *AUC* might be a particularly robust measure for use in BLA, at least in the cases where base rates are likely to vary across samples.

Phase 1 results summary. The findings from the initial restriction of range analyses indicate the each of the accuracy measures (*r*, the *AUC*, and *OR*) were impacted by variations in the range of the predictor variable regardless of which similarity coefficient was used. However, the *AUC* in particular seemed to be affected by this manipulation. The real surprise that was encountered when conducting this analysis was the extreme values for the *OR*s and the very wide CIs associated with these values for both the serial rape and burglary samples. The scores that were obtained for the *OR*s suggest that it is an inappropriate accuracy measure to use in the context of BLA given how similarity coefficients are currently handled by researchers (i.e., ranging as they do from 0 to 1, rather than relying on a scaled version of the coefficients).

For the base rate manipulation, *r* proved to be particularly sensitive (especially in the case of rape) while the *AUC* was generally unaffected. Given that the *AUC* appears to be slightly more impacted by the restriction of range issue than *r*, but *r* seems to be slightly more impacted by the base rate issue compared to the *AUC*, Phase 2 proceeded using both *r* and the *AUC* to determine which similarity coefficient (*J*, *S*, or *SD*) would prove most effective for linking serial rapes and burglaries.

Phase 2

In order to compare the ability of *J, S,* and *SD* to link serial rapists and burglars, the three similarity coefficients for the two samples (rape and burglary) were compared using *r* and the *AUC* as accuracy measures. Both accuracy measures were used for the comparison of similarity coefficients in Phase 2 since, as discussed above, the *AUC* was found to be impacted to a greater degree by the restriction of range manipulations, but *r* was more impacted by the base rate manipulations. Furthermore, to ascertain any effect

that using unequal versus equal samples may have on results, the comparison of similarity coefficients (*J, S,* and *SD*) across the two samples (rape and burglary) for both accuracy measures (*r* and the *AUC*) was carried out twice, once using unequal samples and again using equal samples.

Linking accuracy based on *r*. In order to examine the accuracy of linking decisions using *r*, the samples of serial rape and burglary were first randomly split in half to form development and validation samples. Logistic regression analyses were then conducted on the development samples and applied to the validation samples, for each of the similarity coefficients separately, so that correlations could be calculated between the predicted probabilities of crime pairs being linked and the actual linkage status of those crime pairs (i.e., actually linked or unlinked). Recall that two different sampling procedures were examined here, with separate regression analyses being conducted for each procedure. One of the procedures was to create unequal samples, where the number of linked and unlinked rape/burglary pairs was unequal. The other procedure was to create equal samples, where the number of linked and unlinked rape/burglary pairs was equal.

Results of the regression analysis provided in Table 6, which is based on the development sample, show that linked crime pairs are consistently characterized by higher levels of across-crime similarity compared to unlinked crime pairs, regardless of which similarity coefficient was being examined, which crime type was being considered, or which sampling method was being used. Additionally, the Wald's and chi-square tests confirmed that all regression models accurately predict whether crime pairs are linked or unlinked (all p's $< .001$) regardless of similarity coefficient, crime type, or

sampling method. However, the observed R^2 values indicate that model fit varied depending on the sample.

Table 6

Summary of the Logistic Regression Results

Coefficient	Rape					Burglary				
	B (SE)	Wald (*df*)	χ^2 (*df*)	R^2	*p*-value	*B* (SE)	Wald (*df*)	χ^2 (*df*)	R^2	*p*-value
Unequal										
J	8.41 (0.90)	87.03 (1)	83.29 (1)	.14	.001	2.69 (0.37)	52.31 (1)	47.99 (1)	.03	.001
S	10.82 (1.48)	53.45 (1)	59.00 (1)	.10	.001	4.56 (0.72)	40.49 (1)	42.53 (1)	.02	.001
SD	7.68 (0.90)	73.28 (1)	80.31 (1)	.13	.001	2.28 (0.33)	46.91 (1)	47.46 (1)	.03	.001
Equal										
J	8.60 (1.62)	28.15 (1)	42.40 (1)	.35	.001	1.70 (0.55)	9.62 (1)	10.06 (1)	.03	.002
S	11.96 (2.40)	24.83 (1)	32.83 (1)	.28	.001	4.34 (0.98)	19.59 (1)	20.98 (1)	.06	.001
SD	7.13 (1.33)	28.81 (1)	41.00 (1)	.34	.001	1.43 (0.46)	9.66 (1)	9.96 (1)	.03	.002

Note. SE, standard error; χ^2, model chi-square; *df*, degrees of freedom; R^2, Nagelkerke index.

Two major findings with respect to R^2 values seem worthy of mention. First, for both the equal and unequal samples, the values of R^2 appear to be noticeably higher for rape cases compared to burglary cases. Indeed, the R^2 values for burglary are consistently low (range: .03 to .06), but the range for rape are substantially higher (range: .10 to .35).

Second, in the case of rape, the sample containing an equal number of linked and unlinked crimes is associated with particularly high R^2 values (range: .28 to .35) compared to the unequal sample (range: .10 to .14). Such a difference was not observed across equal and unequal samples in the case of burglary.

The results provided in Table 7 are based on the validation samples. Results show high levels of linking accuracy, with all r's being significant (all p's $< .01$). This was the case for each similarity coefficient that was tested, across the unequal and equal samples, and for each type of crime. However, several interesting results emerged from the analysis.

First, although most of the CIs overlap within each sample, regardless of which similarity coefficient is used, this is not the case for the rape unequal sample where the CI for *S* does not overlap with the CIs for *J* and *SD*. This indicates that *J* and *SD* significantly outperformed *S* in the case of rape when the sample was unequal.

Second, the CIs found for the rape unequal sample do not overlap with the CIs for the rape equal sample. This suggests that linking accuracy, as measured by *r*, may be higher for equal samples of rape compared to unequal samples. The exact same results were found for the serial burglary analysis, with linking accuracy being significantly higher for the equal sample. The fact that differences were found between the equal and unequal samples when using *r* is perhaps not surprising given the previous results from the base rate manipulation, which showed that *r* can be affected by differences in base rates (the base rate of linked crimes is obviously lower in unequal samples).

Table 7

Correlations for Each Regression Model

Coefficient	Rape		Burglary	
	r (CI_{84})	Standard Deviation	r (CI_{84})	Standard Deviation
Unequal				
J	.24 (.23-.25)	.02	.08 (.07-.09)	.02
S	.15 (.14-.17)	.02	.06 (.05-.07)	.02
SD	.24 (.23-.25)	.02	.07 (.06-.08)	.02
Equal				
J	.56 (.49-.62)	.07	.20 (.14-.25)	.25
S	.50 (.43-.56)	.06	.24 (.19-.30)	.25
SD	.55 (.48-.61)	.07	.20 (.14-.25)	.25

Note. Rape unequal sample (n = 126 linked pairs; n = 7749 unlinked pairs); rape equal sample (n = 252); burglary unequal sample (n = 420 linked pairs; n = 21525 unlinked pairs); burglary equal sample (n = 840). All correlations significant ($p < .01$).

Third, the CIs between the serial rape and burglary samples did not overlap for both the unequal and equal samples. The lack of overlap between rape and burglary across both groups (unequal and equal) suggest that linking accuracy was significantly higher in cases of rape compared to burglary when using r as the measure of linking accuracy. Again, this result is unsurprising given previous analyses. For example, the distributions of similarity scores depicted in Figures 1 to 6 clearly indicate less distribution overlap in the case of serial rape, which suggests that it should be easier to distinguish between linked and unlinked rapes than it is to distinguish between linked and unlinked burglaries.

Linking accuracy based on the *AUC*. Using the same logistic regression models described in Table 6, *AUC*s were calculated using data from the validation samples by submitting the predicted probabilities (and linkage status) to ROC analysis. Similar to *r*, comparisons between similarity coefficients, crime type (rape and burglary), and unequal and equal sample sizes were conducted.

The results of the ROC analyses are presented in Table 8, and the ROC curves for each model can be seen in Figures 7 through 18. The *AUC*s from the ROC analyses confirm the logistic regression findings in that all models showed high levels of linking accuracy (all p's < .001). Abiding by the *AUC* guidelines set out by Swets (1988), where low accuracy equals .50 - .70, good accuracy equals .70 - .90, and high accuracy equals > .90, it can be seen that the rape data were associated with good levels of accuracy for both the equal and unequal samples, regardless of which similarity coefficient was used (all CIs overlapping). The results for the burglary data were slightly less impressive with all of the results indicating relatively low levels of accuracy; yet, like the rape data, all CIs overlapped regardless of which similarity coefficient was used. As was the case with *r*, the *AUC*s associated with the rape data were significantly higher than the *AUC*s associated with the burglary data, suggesting that it is possible to discriminate between linked and unlinked crimes to a greater degree in cases of rape compared to burglary. Interestingly, and as should be expected given the previous examination of the base rate issue, the differences in linking accuracy between the equal and unequal samples that were observed when using *r* are not observed here (all CIs overlapping).

Table 8

AUCs for Each Regression Model

	Rape			Burglary		
Coefficient	*AUC*	SE	CI_{84}	*AUC*	SE	CI_{84}
Unequal						
J	.78	.02	.75-.82	.62	.01	.60-.64
S	.77	.02	.74-.80	.62	.01	.60-.64
SD	.78	.02	.75-.82	.62	.01	.60-.64
Equal						
J	.81	.03	.78-.85	.62	.02	.59-.64
S	.79	.03	.75-.83	.64	.02	.61-.67
SD	.81	.03	.78-.85	.62	.02	.59-.64

Note. AUC, area under the curve; SE, standard error; CI_{84}, 84% confidence intervals.

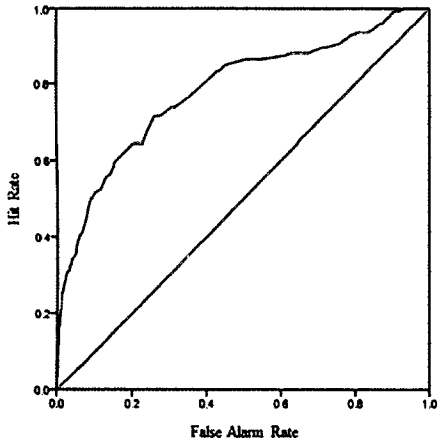

Figure 7. ROC graph for rape unequal model using *J* as the similarity coefficient (*AUC* = .78).

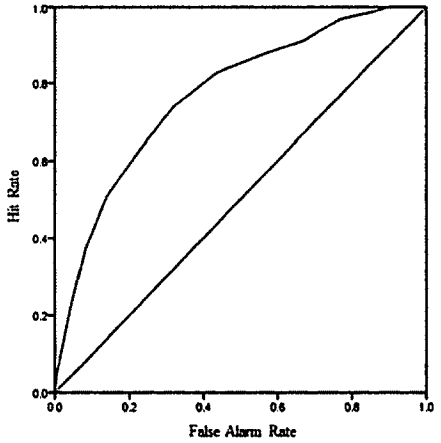

Figure 8. ROC graph for rape unequal model using *S* as the similarity coefficient (*AUC* = .77).

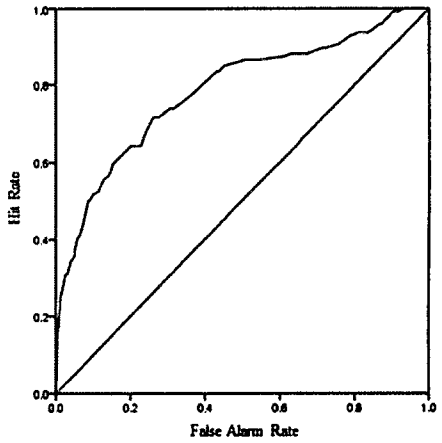

Figure 9. ROC graph for rape unequal model using *SD* as the similarity coefficient (*AUC* = .78).

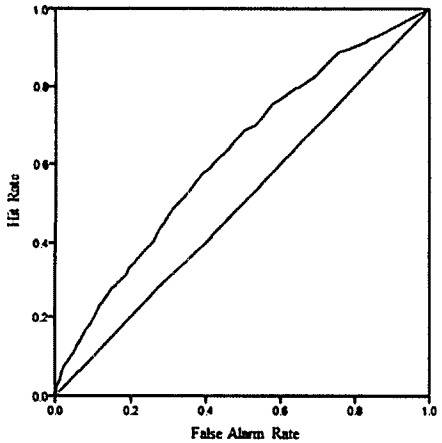

Figure 10. ROC graph for burglary unequal model using *J* as the similarity coefficient (*AUC* = .62).

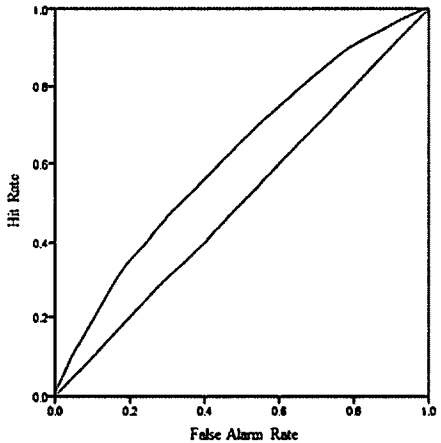

Figure 11. ROC graph for burglary unequal model using *S* as the similarity coefficient (*AUC* = .62)

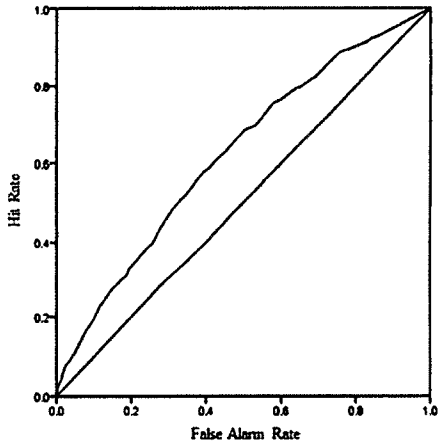

Figure 12. ROC graph for burglary unequal model using *SD* as the similarity coefficient (*AUC* = .62).

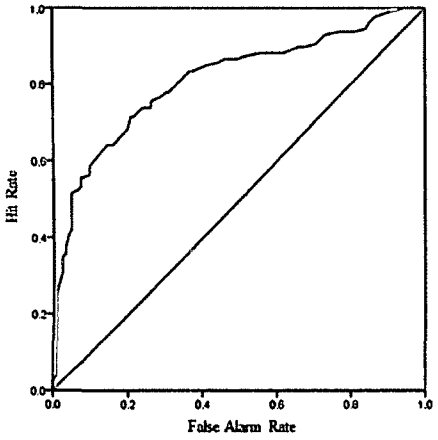

Figure 13. ROC graph for rape equal model using *J* as the similarity coefficient (*AUC* = .81).

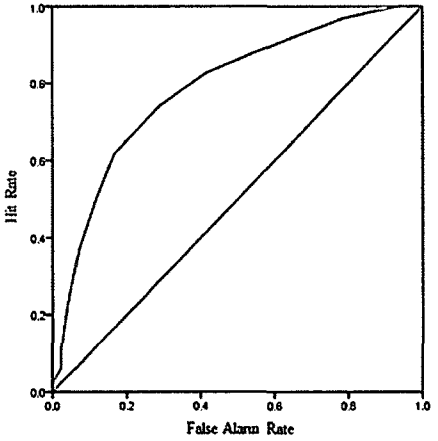

Figure 14. ROC graph for rape equal model using *S* as the similarity coefficient (*AUC* = .79).

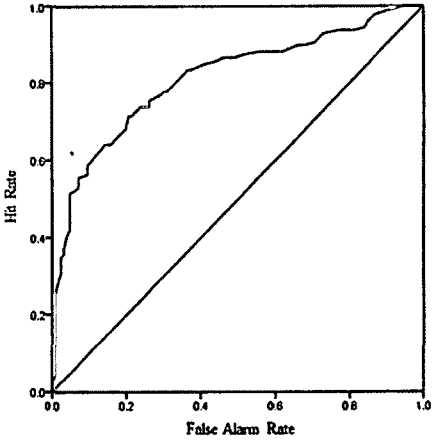

Figure 15. ROC graph for rape equal model using *SD* as the similarity coefficient (*AUC* = .81).

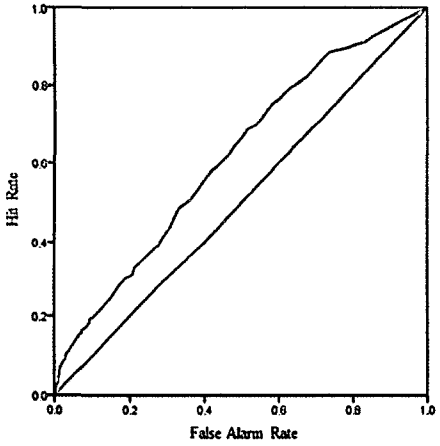

Figure 16. ROC graph for burglary equal model using *J* as the similarity coefficient (*AUC* = .62).

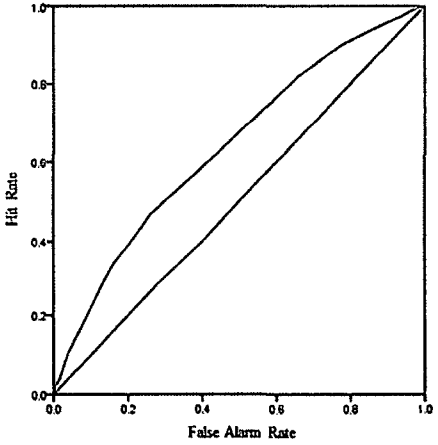

Figure 17. ROC graph for burglary equal model using *S* as the similarity coefficient (*AUC* = .64).

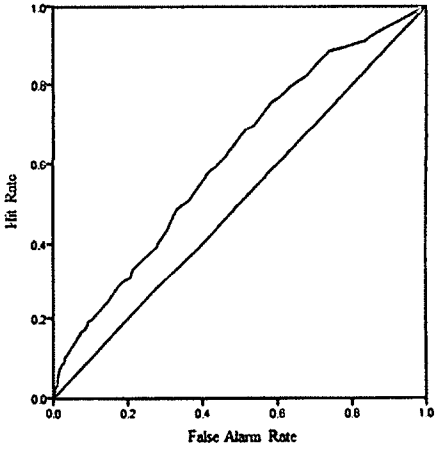

Figure 18. ROC graph for burglary equal model using *SD* as the similarity coefficient (*AUC* = .62).

Discussion

The main purpose of this study was to expand existing research on BLA by examining two key issues. First, three linking accuracy measures - *r*, the *AUC*, and the *OR* - were compared across two manipulations (restriction of range in the predictor variables and base rate variability) to determine which measure is the best suited (i.e., most robust) for BLA research. To increase the generalizability of the results, different crime types (serial rape and serial burglary) were used for this comparison as well as three different similarity coefficients (*J*, *S*, and *SD*).

Second, using the most robust measure(s) of linking accuracy from the first phase of the research, the goal was to determine which similarity coefficient (*J*, *S*, or *SD*) was associated with the highest levels of linking accuracy when predicting crime linkages in cases of serial rape and burglary. In order to resolve an ongoing debate in the BLA literature about how samples should be constructed, all of the analyses in Phase 2 were conducted using samples of linked and unlinked crime pairs that were either unequal or equal in number, the former procedure being more similar to what occurs in naturalistic settings.

At the outset it was felt that the results of this study have the potential to add to our existing theoretical understanding of BLA, with respect to the consistency and distinctiveness of criminal behaviour, and to address key practical questions, such as the measure of linking accuracy that should be focused on in BLA research, the similarity coefficient that should be used, and the best way to sample crimes. These findings can better inform us as to which conditions provide the greatest discrimination accuracy for

behavioural linking. The implications of this study's findings will now be discussed in more detail below.

Evidence for Behavioural Consistency and Distinctiveness

Prior to examining the two issues described above, it was first deemed important to examine the degree to which behavioural consistency and distinctiveness exist within the samples of serial rapists and burglars that are the focus of the current study. Descriptive statistics of across-crime similarity scores for both linked and unlinked crimes were calculated for this purpose, and the distributions of these scores were plotted. Simple inferential tests were also applied to these values to identify if any significant differences existed between the scores for linked and unlinked crime pairs.

This initial set of analyses indicated that the sampled offenders do behave in a relatively stable and distinct fashion, which should allow for a reasonable level of discrimination accuracy to be achieved in BLA. For example, across-crime similarity scores were consistently higher for linked compared to unlinked crimes, for both serial rapists and burglars. This finding supports previous research, which has indicated that offenders behave in a relatively stable and distinct fashion across their crimes (e.g., Goodwill & Alison, 2006; Grubin et al., 2001; Markson, Woodhams, & Bond, 2010). It also suggests that serial offenders, much like their non-criminal counterparts, possess predispositions to behave in particular ways (e.g., Greene, 1989; Meyer, 1990; Mischel, Shoda, & Mendoza-Denton, 2002; Shoda, 1999).

Interestingly, the results from this initial analysis also suggested that the ability to discriminate between linked and unlinked crimes will likely be higher for rape than burglary. For example, the distribution of similarity scores for linked and unlinked crimes

overlapped to a lesser degree for rape compared to burglary. When considering the sort of behaviors relied upon in this study, this is consistent with previous research that has also found that linking accuracy tends to be lower for property crimes than interpersonal crimes (Markson et al., 2010; Tonkin et al., 2008; Woodhams & Labuschagne, 2011; Woodhams & Toye, 2007). In the few instances where high levels of linking accuracy are found for crimes like burglary, very specific subsets of behaviours are examined, often related to crime site selection choices (e.g., Bennell & Canter, 2002; Bennell & Jones, 2005).

Why might serial rapists exhibit higher levels of stability and distinctiveness than serial burglars? There are many potential answers to this question, but one possibility is that, unlike burglars, rapists often rely upon personal scripts when offending; scripts that are sometimes based on deeply ingrained fantasies involving behaviours that are often well rehearsed before the crime is committed (Davies, 1992; Douglas & Munn, 1992; Geberth, 1996; Hazelwood & Warren, 2003; Hickey, 1991; Keppel, 1997). These scripts could arguably lead to relatively unique, stable patterns of behaviour. Another, more practical explanation, may relate to the reliability of the data that researchers have to work with when examining BLA. More specifically, data from crimes like rape may be more reliable, and thus more likely to reveal patterns of behavioural stability and distinctiveness, compared to burglary data. This is likely the case because investigating officers must often make inferences when reconstructing burglaries of unoccupied residences (e.g., when a crime occurred, how entry to the property was gained, what valuable items were ignored, what internal behaviours were exhibited, when the burglary actually took place, etc.; e.g., Ratcliffe, 2002), whereas victims are present in cases of

rape and can report on the behaviour of offenders when providing their statements to the police.

Which Accuracy Measure is the Most Effective for BLA?

Recall that a number of accuracy measures can be used to assess BLA and therefore questions are often raised about which measure should be used by researchers. There are obviously a variety of ways to assess the suitability of linking accuracy measures, but part of the assessment should arguably seek to understand the extent to which the measures under examination are robust (e.g., unaffected by factors that vary naturally in the BLA context). For example, we might want to use a measure of accuracy that does not vary as a function of how restricted the across-crime similarity scores are that form the basis of BLA, and presumably we would want to use an accuracy measure that is not heavily influenced by fluctuations in the base rate within the samples being analyzed (i.e., the number of linked crime pairs). In an attempt to determine whether some accuracy measures are more robust than others across these conditions, we attempted to replicate the study conducted by Hanson (2008) in the context of BLA research.

In his study, Hanson (2008) manipulated two factors (restriction of range in the predictor variable and variability in the base rates) and found that *r* and the *AUC* were affected by variability in restriction of range, whereas only *r* was affected by variability in the base rates. Despite some previous research suggesting that the *OR* can also be affected by base rate variability (Rice & Harris, 1995), Hanson found that the *OR* was unaffected by either manipulation. This led him to conclude that the *OR* may be a particularly useful measure of predictive accuracy, at least in the context of risk

assessment decisions for sex offenders. Based on Hanson's study, it was predicted that *r* and the *AUC* would prove sensitive to restriction of range in the BLA context, but that only *r* would be affected by variability in the base rates. The *OR* was expected to be unfazed by either manipulation.

What is the influence of restriction of range? When restriction of range was manipulated for the serial rape and burglary data, the hypotheses were only partially supported. For example, across both datasets (rape and burglary) there seemed to be a similar degree of variability (with respect to non-overlapping CIs) for *r* and the *AUC*. This was consistent with Hanson's (2008) results and our expectations. However, in contrast to Hanson's findings, it was found that only the *AUC* showed strong correlations between the variability in across-crime similarity scores and the observed magnitude of the *AUC*s (i.e., for rape: *J*, $r = .86$; *SD*, $r = .73$; for burglary: *J*, $r = .71$; *SD*, $r = .94$). No significant correlations were found for *r*. Interestingly, significant correlations between the *AUC* values and variability in the similarity scores were not observed for every similarity coefficient that was tested, only for *J* and *SD*.

Surprisingly, the correlation that was based on *S* for the *AUC* did not reach statistical significance for either sample (rape or burglary). The most likely explanation for this relates to the fact that there were no *S* scores in the low sub-samples for rape and burglary and no linked *S* scores for moderately low sub-samples for rape. As indicated in the initial descriptive analysis, this is because *S* includes joint non-occurrences in its calculation, which leads to relatively high across-crime similarity scores (for both linked and unlinked crimes) compared to *J* and *SD*. Given the more restricted range associated

with *S* scores, it is arguably not well-suited to studying the robustness of accuracy measures to manipulations of range restriction.

Consistent with our expectations and Hanson's (2008) study, the CIs associated with the *OR*s largely overlapped and non-significant correlations were found between the variability in the similarity scores and the observed magnitude of the *OR*s, except in the cases of *J* and *S* for the rape sample. However, as mentioned already, these results were difficult to interpret given the extreme scores (and very wide CIs) that were obtained. These extreme values were not expected and likely came about because of the relatively unusual nature of the predictor variables used in studies of BLA; that is, that the range of the similarity scores is a maximum of 1. Indeed, for most of the sub-samples tested in the analysis of the range restriction manipulation, the range of similarity scores is <1, making it unclear what the *OR*s (and their CIs) even mean (Craig Leth-Steensen, personal communication, August 27, 2012).

In contrast to the conclusion reached by Hanson (2008), which was that the *OR* is a particularly robust (and therefore useful) measure of predictive accuracy, the results presented in the current paper suggest that the value of the *OR* may depend on study-specific factors (e.g., the maximum range of predictor variables). This conclusion is consistent with arguments made by others that the suitability of the *OR* depends on the task under consideration (Rice & Harris, 1995; Rosenthal, 1991). Of course, the results of the current study do not necessarily mean that the *OR* might not be a useful accuracy measure for BLA research per se, just that it is a problematic measure given how BLA is currently conducted. It would be useful in the future to scale the similarity scores tested in BLA research by, for example, multiplying the scores by a value of 100. This would

eliminate the problems encountered by having to rely on predictor variables with a range <1. When this was explored in the current study (see footnote 6), the *r* and *AUC* values remained unchanged, but the *OR*s became more interpretable. Under these conditions, the *OR* may prove to be a robust and useful measure for BLA research.

What is the influence of base rate manipulation? Given the findings for the *OR*s in the analysis of range restriction, only *r* and the *AUC* were examined in the analysis of the base rate manipulation. For *r*, as the base rate (i.e., number of linked crime pairs) decreased, so too did the magnitude of the effect size, but in the current study, this did not generally occur to a statistically significant degree. The *AUC* was unaffected by base rate fluctuations. These results were the same regardless of crime type or similarity coefficient used, suggesting some level of generalizability across conditions. Although we expected significant variability in *r* for the base rate manipulation, at a descriptive level at least all of these findings are generally consistent with the findings of Hanson (2008) and others (e.g., Rice & Harris, 1995), where *r* was found to be more sensitive to base rate fluctuations than the *AUC*.

Of course there are other advantages associated with the *AUC* beyond the fact that it is resistant to fluctuations in base rate (Bennell & Jones, 2005). For example, unlike *r*, the *AUC* is completely independent of the decision threshold (i.e., similarity score) used to make linking decisions. Indeed, because the *AUC* represents the area under the entire ROC curve on a ROC graph, which is made up of hit to false alarm ratios for multiple decision thresholds, the *AUC* is not dependent on any one of these ratios. Given these advantages, it may be that the *AUC*, as a measure of predictive accuracy, is better suited to BLA research. However, given the results reported in this study, where potential

problems emerged for both *r* and the *AUC*, a decision was made to explore each of these measures in Phase 2 of the study.

Which Similarity Coefficient is Most Appropriate for BLA?

Phase 2 of the current study focused on the BLA task and explored the impact of using different similarity coefficients and sampling procedures. Both *r* and the *AUC* were used as measures of linking accuracy.

Which similarity coefficient was associated with the highest level of accuracy? There was one instance where *S* performed significantly worse than *J* and *SD*, which was for the unequal rape sample when using *r* as the measure of accuracy. Otherwise, although *S* performed slightly worse than *J* and *SD* in all cases, all similarity measures performed at around the same level for both rape and burglary. This is somewhat inconsistent with the study conducted by Ellingwood et al. (in press), where the use of *S*, on one occasion at least, resulted in higher levels of linking accuracy than *J*.

The fact that very few significant differences were found across the different similarity coefficients is very interesting, especially given the fact that some incorporate information about joint non-occurrences, whereas others do not. This result is also interesting when one considers the sorts of distributions associated with these different similarity coefficients. For example, due to its incorporation of joint non-occurrences, similarity scores for *S* (for both linked and unlinked crimes) are larger than the similarity scores for *J* and *SD* (i.e., the distributions of scores fall higher on the x-axis in Figures 2 and 5). Clearly, what matters in the context of BLA then, is not where these distributions fall along the x-axis, but the degree to which the linked and unlinked distributions

overlap (Bennell, Jones, & Melnyk, 2009).[9] Given the results that were found across the range restriction and base rate manipulations, it appears that this degree of overlap is similar across the three coefficients, and thus they all appear to be equally well-suited to the BLA task (with the exception perhaps of unequal rape samples).

What is the impact of crime type? Although the different similarity coefficients performed equally well, with the exception of that one instance where *S* performed significantly worse than *J* and *SD*, linking accuracy was found to vary as a function of crime type. Specifically, the level of accuracy found in Phase 2 of the current study was significantly lower in the case of serial burglary compared to rape for all three similarity coefficients, regardless of whether *r* or the *AUC* was used to measure linking accuracy. Possible reasons for this difference have already been discussed (e.g., the reliance on deeply engrained personal scripts in cases of rape and higher levels of reliability that might be associated with rape data), however there may also be another factor that accounts for this difference - the type and number of behaviours included in the analysis.

As discussed above, serial burglars tend to only show high levels of behavioural stability and distinctiveness when very specific sub-sets of behaviours are examined. More specifically, stability and distinctiveness are found for inter-crime distances and temporal proximity, but rarely for the sorts of behaviours examined in the current study (information related to crime site locations and dates of occurrence were not available in the current study) (Markson et al., 2010). On the other hand, high levels of linking accuracy has been found for interpersonal crimes when using more traditional MO

[9] As argued elsewhere (e.g., Bennell & Canter, 2002), the degree of overlap between linked and unlinked distributions reflect how ambiguous across-crime similarity scores are. If the distributions overlap completely, this means that any specific similarity score, regardless of its value, is just as likely to be associated with linked crimes as they are to be associated with unlinked crimes.

indicators, such as methods used by the offender to control the victim (Woodhams & Toye, 2007). Thus, unlike the situation for burglary, the data that was available in the serial rape dataset within the current study was likely to reveal patterns of behavioural stability and distinctiveness.

It is also important to highlight the fact that the number of crime scene behaviours available for analysis differed across the rape and burglary datasets. In the case of rape, 36 behaviours were available, whereas only 27 behaviours were available in the case of burglary. It is possible that including more behaviours in BLA results in higher levels of linking accuracy. This was the case in Ellingwood et al.'s (in press) recent examination of BLA in serial arson cases. They found that linking accuracy was higher when using larger number of arson behaviours than when using sub-sets of arson behaviours. Presumably this is because the use of more behaviours allows one to account for more variance in the criterion variable. However, little research has examined this particular issue and thus it is too early to draw any strong conclusions. Future research should examine this issue further.

What is the Influence of Sampling Procedure? Another goal of Phase 2 was to see how linking accuracy might be affected by the sampling procedure used by the researcher, specifically whether unequal or equal samples are relied on (i.e., whether the number of linked and unlinked crime pairs is unequal or equal). The primary debate over using unequal or equal samples is one of ecological validity versus statistical reliability. The reliance on unequal samples is more ecologically valid since investigators will be exposed to all linked and unlinked crime pairs when carrying out their analysis. However, an unequal sample might bias the results of BLA given the disproportionate number of

unlinked crime pairs in the sample. The results of the current study revealed some intriguing findings regarding this debate.

In the current study, the *AUC* results suggest that the choice of unequal or equal sample sizes had little impact on the linking results, regardless of crime type or similarity coefficient. However, the same cannot be said when *r* was used. In this case, significant differences were found between the unequal and equal sample sizes, regardless of crime type or similarity coefficient. More specifically, based on the non-overlapping CIs, *r* was consistently higher when equal samples were used. This is not particularly surprising considering that Phase 1 of the study determined that *r* was slightly more sensitive to base rates than the *AUC*, which is something that varies dramatically across equal and unequal samples (the base rate of linked crimes being much higher in the case of equal samples). Again, this finding potentially speaks to the value of using the *AUC* as the preferred measure of accuracy for BLA research.

Study Limitations

The current study was limited in a number of ways. First, the crimes included in the study were all solved crimes and crimes might be solved because the offenders displayed higher levels of stability and distinctiveness than crimes that are not solved (Bennell & Canter, 2002). If this is true, then the linking accuracy levels reported in this study may be an overestimate of how accurate crime linkages would be in naturalistic settings where BLA is obviously applied to unsolved crimes. This limitation is inherent to virtually all linking research, although very recent efforts have been made to get around this problem by relying of serial crimes that were first linked by DNA rather than offenders' MO (Woodhams & Labuschagne, 2012).

Second, only serial offenders were used in the current study. In real life investigative situations, both serial and non-serial offences would be included in linking analysis. Of course, the inclusion of one-off offenders would make the linking task far more difficult for investigators, but such samples would better reflect the reality that crime analysts face when conducting BLA. The exclusion of one-off studies in the current study (and in other linking studies) would presumably lead to overestimates of linking accuracy. To go beyond current research, future studies of BLA should include one-off criminals so that the levels of linking accuracy emerging from these studies are more likely to generalize to naturalistic settings.

Third, there may be issues with the reliability of the data used in the current study given that the raw data was not collected for the purpose of a research study. In the case of rape, the data was based on victim statements and therefore there are a number of issues that we should be concerned with. For example, victims might emphasize certain behaviours over others, they may not remember the occurrence of certain behaviours, or they may be too embarrassed to report them to the police (Alison, Snook, & Stein, 2001). In the case of residential burglaries where witnesses are rare and victims are not present during the commission of the crime, police officers must often make inferences about what the offender did and crime reports are often filled with best guesses. While the presence of a victim in rape cases may mean that this data is slightly more reliable than burglary data, clearly we should be cautious of all the data used in the current study given these issues.

Fourth, and finally, capping the number of crimes included in each offender's series, as was done in the current study, is a potential limitation. As with other studies

that have taken this approach (e.g., Markson et al., 2010; Tonkin et al., 2008; Woodhams, Grant, et al., 2007; Woodhams & Toye, 2007), this was done here to limit the potentially biasing effect of having prolific offenders in the sample that are particularly stable and/or distinct (or lacking stability and/or distinctiveness). While this sampling approach accomplishes this, it does not reflect the situation faced by investigators when conducting BLA. In the real world, investigations include all offences in a crime series and by sampling crimes in the way we did here, the external validity of the study might be compromised.

Future Directions

In addition to the suggestions for future research that have already been made, there are a variety of other issues that could be explored in future research. For example, there are a wide variety of similarity coefficients that can be examined beyond the three that were examined in the current study. Without exhausting the range of similarity coefficients that could be tested, it is not possible to draw conclusions about which similarity coefficient should be used for BLA, if indeed there is one measure that will lead to optimal results.

Additional research using both unequal and equal samples would also be worthwhile to ensure that the results of this study generalize to other samples. Examining a broader range of crime types would also be useful, especially to see if the differences in linking accuracy that were found between the serial rape and burglary samples could be replicated with other interpersonal and property crimes. Using crime data from other countries should also be a priority to ensure that the results of the current study are not

biased by cultural factors. Including one-off crimes in future research would also be helpful and would make that research more ecologically valid.

Finally, further tests of the *OR* are warranted to see if the problems encountered in this study hold up in future research. If they do, this is important information for BLA researchers to know. Clearly, one direction this research should take is to examine more thoroughly the impact that scaling similarity scores has on *OR*s. It is possible that this future research will confirm the conclusions of Hanson (2008), and support the use of *OR*s in the BLA context. If, after these future studies, the *AUC* is still the recommended measure of linking accuracy for BLA research, it will be important to examine other factors, beyond range restriction and base rate variability, which may influence this measure (e.g., when large amounts of data is missing, as is often the case when relying on victim statements and crime reports).

Conclusion

The current study sought to answer two important questions: which accuracy measure is best suited for BLA research (*r*, the *AUC*, or the *OR*) and which similarity coefficient results in optimal linking accuracy. In contrast to previous studies, the *OR* has been identified as a potentially inappropriate measure for BLA research, at least as it is currently being conducted. The types of similarity coefficients that are currently used, or more specifically the range of similarity scores that results from the use of these coefficients, appears to result in *OR*s that are very difficult to interpret (and are potentially not meaningful). Instead, *r* and the *AUC* are more appropriate choices as measures of linking accuracy, though the results from Phase 2 suggest that when there are base rate differences between samples (e.g., as was the case when using equal and

unequal samples) the *AUC* might be the most suitable measure. With one exception, where *S* resulted in significantly lower levels of linking accuracy than *J* and *SD*, the similarity coefficients that were tested in this study produced similar results. Clearly, future research examining other similarity coefficients would be useful, but the current results suggest that each of the coefficients examined in this study may all be suitable for use in BLA as it is currently studied.

References

Alison, L. J., & Stein, K. L. (2001). Vicious circles: Accounts of stranger sexual assault reflect abusive variants of conventional interactions. *Journal of Forensic Psychiatry, 12*, 515-538. doi: 10.1080/09585180127391

Alison, L. J., Snook, B., & Stein, K. L. (2001). Unobtrusive measurement: Using police information for forensic research. *Qualitative Research, 1*, 241-254. doi: 10.1177/146879410100100208

Bennell, C. (2002). *Behavioural consistency and discrimination in serial burglary.* Unpublished doctoral dissertation, University of Liverpool, Liverpool, UK.

Bennell, C., Bloomfield, S., Snook, B., Taylor, P. J., & Barnes, C. (2010). Linkage analysis in cases of serial burglary: comparing the performance of university students, police professionals, and a logistic regression model. *Behavioral Sciences and the Law, 16,* 507-524. doi: 10.1080/10683160902971030

Bennell, C., & Canter, D. V. (2002). Linking commercial burglaries by modus operandi: Tests using regression and ROC analysis. *Science & Justice, 42*, 153-164. doi:10.1016/S1355-0306(02)71820-0

Bennell, C., Gauthier, D., Gauthier, D., Melnyk, T., & Musolino, E. (2010). The impact of data degradation and sample size on the performance of two similarity coefficients used in behavioural linkage analysis. *Forensic Science International, 199*, 85-92. doi: 10.1016/j.forsciint.2010.03.017

Bennell, C., & Jones, N. J. (2005). Between a ROC and a hard place: A method for linking serial burglaries by modus operandi. *Journal of Investigative Psychology and Offender Profiling, 2,* 23-41. doi: 10.1002/jip.21

Bennell, C., Jones, N. J., & Melnyk, T. (2009). Addressing problems with traditional crime linking methods using receiver operating characteristic analysis. *Legal and Criminological Psychology, 14*, 293-310. doi: 10.1348/135532508X349336

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods, 14*(3), 225-238. doi: 10.1037/a0016619

Borenstein, M., Hedges, L., & Rothstein, H. R. (2007). Introduction to Meta-Analysis. In *meta-analysis.com*. Retrieved April 16, 2012, from http://www.meta-analysis.com/downloads/Meta%20Analysis%20Fixed%20vs%20Random%20effects.pdf.

Canter, D. V. (1995). Psychology of offender profiling. In R. Bull & D. Carson (Eds.), *Handbook of psychology in legal contexts* (pp. 343-355). Chichester, UK: Wiley.

Canter, D. V., Bennell, C., Alison, L., & Reddy, S. (2003). Differentiating sex offences: A behaviorally based thematic classification of stranger rapes. *Behavioral Sciences & the Law, 21*, 157-174. doi: 10.1002/bsl.526

Canter, D. V., & Fritzon, K. (1998). Differentiating arsonists: A model of firesetting actions and characteristics. *Legal and Criminological Psychology, 3,* 73-96. doi: 10.1111/j.2044-8333.1998.tb00352.x

Canter, D. V., & Larkin, P. (1993). The environmental range of serial rapists. *Journal of Environmental Psychology, 13*(1), 63-69. doi: 10.1016/S0272-4944(05)80215-4

Canter, D. V., Wilson, M., Jack, K., & Butterworth, D. (1996). *The psychology of rape investigations: A study in police decision making.* Liverpool, UK: University of Liverpool.

Cervone, D., & Shoda, Y. (1999). Beyond traits in the study of personality coherence. *Current Directions in Psychological Science, 8*, 27-32. doi: 10.1111/1467-8721.00007

Dalirsefat, S., Meyer, A., & Mirhoseini, S. (2009). Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the silkworm, *Bombyx mori. Journal of Insect Science, 9*(71), 1-8.

Davies, A. (1992). Rapists' behaviour: A three aspect model as a basis for analysis and identification of serial crime. *Forensic Science International, 55*, 173-194. doi: 10.1016/0379-0738(92)90122-D

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*, 297-302. doi: 10.2307/1932409

Douglas, J. E., Burgess, A. W., Burgess, A. G., & Ressler, R. K. (1992). *Crime classification manual.* New York: Simon & Schuster.

Douglas, J. E., & Munn, C. (1992). Violent crime scene analysis: Modus operandi, signature, and staging. *FBI Law Enforcement Bulletin, 61*, 1-10.

Efron, B. (1982). *The jackknife, the bootstrap, and other re-sampling plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Ellingwood, H., Mugford, R., Melnyk, T., & Bennell, C. (in press). Linking serial arson: Comparing the Simple Matching Index to Jaccard's Coefficient. *Journal of Investigative Psychology and Offender Profiling*. doi: 10.1002/jip.1364

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). California, USA: SAGE Publications.

Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology, 60*, 773–794. doi: 10.1037/0022-3514.60.5.773

Furr, R. M., & Funder, D. C. (2003). Situational similarity and behavioural consistency: Subjective, objective, variable-centred and person-centred approaches. *Journal of Research in Personality, 38*, 421-447. doi: 10.1016/j.jrp.2003.10.001

Geberth, V. J. (1996). *Practical homicide investigation* (3rd ed.). Boca Raton, FL: CRC Press.

Goodwill, A.M. & Alison, L.J. (2006). The development of a filter model for prioritising suspects in burglary offences. *Psychology, Crime & Law, 12*, 395-416. doi: 10.1080/10683160500056945

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification, 3*, 5-48. doi: 10.1007/BF01896809

Greene, J. O. (1989). The stability of nonverbal behaviour: An action-production approach to problems of cross-situational consistency and discriminativeness. *Journal of Language and Social Psychology, 8*, 193-220. doi: 10.1177/0261927X8983003

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, & Law, 2*, 293-323. doi: 10.1037/1076-8971.2.2.293

Grubin, D., Kelly, P., & Brunsdon, C. (2001). *Linking serious sexual assaults through behaviour*. London, UK: Home Office.

Häkkänen, H., Lindlöf, P., & Santtila, P. (2004). Crime scene actions and offender characteristics in a sample of Finnish stranger rapes. *Journal of Investigative Psychology and Offender Profiling, 1*, 17-32. doi: 10.1002/jip.1

Häkkänen, H., Puolakka, P., & Santtila, P. (2004). Crime scene actions and offender characteristics in arsons. *Legal and Criminological Psychology, 9*, 197-214. doi: 10.1348/1355325041719392

Hanson, K. (2008). What statistics should we use to report predictive accuracy? *Crime Scene, 15* (1), 15-17.

Harris, A., Phenix, A., Hanson, R. K., & Thornton, D. (2003). *Static-99 coding rules: Revised 2003*. Ottawa: Department of the Solicitor General of Canada.

Hazelwood, R. R., & Warren, J. (2003). Linkage analysis: Modus operandi, ritual, and signature in serial sexual crime. *Aggression and Violent Behaviour, 8*, 587–598. doi: 10.1016/S1359-1789(02)00106-4

Hickey, E. W. (1991). *Serial murderers and their victims*. California, USA: Wadsworth Publishing Company.

Jaccard, P. (1908). Nouvelle recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles, 44*, 223-270.

Keppel, R. D., & Birnes, W. J. (1997). *Signature killers*. New York, N.Y.: Pocket Books.

Kocsis, R. N., & Irwin, H. J. (1997). An analysis of spatial patterns on Australian offences of serial rape, arson and burglary: The utility of the circle theory of environmental range for psychological profiling. *Psychiatry, Psychology and Law, 4*, 195–206. doi: 10.1080/13218719709524910

Kosman, E., & Leonard, K. J. (2005). Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Molecular Ecology, 14,* 415-424. doi: 10.1111/j.1365-294X.2005.02416.x

Labuschagne, G. N. (2010). The use of linkage analysis as an investigative tool and evidential material in serial offenses. In K. Borgeson & K. Kuehnle (Eds), *Serial offenders: theory and practice* (pp. 187-215). Sudbury: Jones & Bartlett Learning.

Liebetrau, A. M. (1983). *Measure of association.* Beverly Hills, CA: Sage Publications.

Markson, L., Woodhams, J., & Bond, J. W. (2010). Linking serial residential burglary: Comparing the utility of modus operandi behaviours, geographic proximity, and temporal proximity. *Journal of Investigative Psychology and Offender Profiling, 7,* 91-107. doi: 10.1002/jip.120

Melnyk, T. (2008). *Factors that influence the accuracy of behavioural linkage analysis in cases of serial sexual assault, homicide, and burglary. Unpublished masters thesis, Carleton University, Ottawa, Canada.*

Melnyk, T., Bennell, C., Gauthier, D., & Gauthier, D. (2011). Another look at across-crime similarity coefficients for use in behavioural linkage analysis: An attempt to replicate Woodhams, Grant, and Price (2007). *Psychology, Crime & Law, 17,* 359-380. doi: 10.1080/10683160903273188

Meyer, J. R. (1990). Cognitive processes underlying the retrieval of compliance-gaining strategies: An implicit rules model. In J. P. Dillard (Ed.), *Seeking compliance: The production of interpersonal influence messages* (pp. 57-74). Scottsdale, AZ: Gorsuch Scarisbrick.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

Mischel, W., Shoda, Y. & Mendoza-Denton, R. (1992). Situation-behaviour profiles as a locus of consistency in personality. *Current Directions in Psychological Science, 11*, 50-54. doi: 10.1111/1467-8721.00166

Mokros, A., & Alison, L. J. (2002). Is offender profiling possible? Testing the predicted homology of crime scene actions and background characteristics in a sample of rapists. *Journal of Legal and Criminal Psychology, 7*, 25–43. doi: 10.1348/135532502168360

Payton, M. E., Greenstone, M. H., & Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science*, *3*(34), 1-6. PMid: 15841220

Ratcliffe, J. H. (2002). Aoristic signatures and the spatio-temporal analysis of high volume crime patterns. *Journal of Quantitative Criminology*, *18*(1), 23-43. doi: 10.1023/A:1013240828824

Ressler, R. K., Burgess, A., & Douglas, J. E. (1988). *Sexual homicide: Patterns and motives.* New York: Lexington Books.

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology*, *63*, 737-748. doi: 10.1037/0022-006X.63.5.737

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. *American Psychologist, 46,* 1086-1087. doi: 10.1037/0003-066X.46.10.1086

Santtila, P., Fritzon, K., & Tamelander, A. L. (2004). Linking arson incidents on the basis of crime scene behavior. *Journal of Police & Criminal Psychology, 19*, 1-16. doi: 10.1007/BF02802570

Santtila, P., Junkkila, J., & Sandnabba, N. K. (2005). Behavioural linking of stranger rapes. *Journal of Investigative Psychology and Offender Profiling, 2*(2), 87–103. doi: 10.1002/jip.26

Santtila, P., Pakkanen, T., Zappalà, A., Bosco, D., Valkama, M., & Mokros, A. (2008). Behavioral crime linking in serial homicide. *Psychology Crime and Law, 14,* 245–265. doi: 10.1080/10683160701739679

Sesli, M., & Yegenoglu, E. D. (2010). Genetic dissimilarities between wild olives by random amplified polymorphic DNA (RAPD) assay. *African Journal of Biotechnology, 9*(53), 8970-8976.

Shoda, Y. (1999). A unified framework for the study of behavioural consistency: Bridging person x situation interaction and the consistency paradox. *European Journal of Personality, 13*, 361-387. doi: 10.1002/(SICI)1099-0984(199909/10)13:5<361::AID-PER362>3.0.CO;2-X

Snook, B., Zito, M., Bennell, C., & Taylor, P. J. (2005). On the complexity and accuracy of geographic profiling strategies. *Journal of Quantitative Criminology, 21(1)*, 1-26. doi: 10.1007/s10940-004-1785-4

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *The University of Kansas Scientific Bulletin, 38,* 1409-1438.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Videnski Selskab Biologiske Skrifter, 5*, 1-34.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285-1293. doi: 10.1126/science.3287615

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics.* New York: Pearson Education.

Tonkin, M., Grant, T. D., & Bond, J. W. (2008). To link or not to link: A test of the case linkage principles using serial car theft data. *Journal of Investigative Psychology and Offender Profiling, 5,* 59-77. doi: 10.1002/jip.74

Tyron, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Journal of Psychological Methods*, *6*(4), 371-386. doi: 10.1037/1082-989X.6.4.37110.1037/1082-989X.6.4.371

Woodhams, J., Grant, T., & Price, A. (2007). From marine ecology to crime analysis: Improving the detection of serial sexual offences using a taxonomic similarity measure. *Journal of Investigative Psychology and Offender Profiling, 4*, 17-27. doi: 10.1002/jip

Woodhams, J., Hollin, C. R., & Bull, R. (2007). The psychology of linking crimes: A review of the evidence. *Legal and Criminological Psychology, 12*, 233-249. doi: 10.1348/135532506X118631

Woodhams, J., & Labuschagne, G. (2011). A test of case linkage principles with solved and unsolved serial rapes. *Journal of Police and Criminal Psychology.* Advance online publication. doi: 10.1007/s11896-011-9091-1

Woodhams, J., & Toye, K. (2007). An empirical test of the assumptions of case linkage and offender profiling with serial commercial robberies. *Psychology, Public Policy, and Law, 13*, 59-85. doi: 10.1037/1076-8971.13.1.59

Wright, R., Decker, S. H., Redfern, A. K., & Smith, D. L. (1992). A snowball's chance in hell: Doing fieldwork with active residential burglars. *Journal of Research in Crime and Delinquency, 29*(2), 148-161. doi: 10.1177/0022427892029002003

Yokota, K., Fujita, G., Watanabe, K., Yoshimoto, K., & Wachi. T. (2007). Application of the behavioral investigative support system for profiling perpetrators of serial sexual assaults. *Behavioural Sciences and the Law, 25*, 841–856. doi: 10.1002/bsl.79

Appendix A

Content Dictionary for Serial Burglary Crime Scene Behaviours (Bennell, 2002)

Twenty-eight variables were created from a content analysis of victim statements in order to provide a list of elements common to offences. All variables are dichotomous with values based on the presence (1) or absence (0) of each category of behaviour. The frequency for each behaviour is given in percentages following the variable name. A description of the categorization scheme in alphabetical order is given below.

1. Alarm (7.6%). The dwelling was equipped with an operational alarm security system at the time of the burglary.

2. Brick (4.3%). Entry was gained by use of brick or stone.

3. Carrier taken (11.4%). The offender took an item from the dwelling in order to carry other items stolen (e.g., pillow case, bag, hold all, etc.).

4. Crowbar (1.9%). Offender gained access to area of crime by using a crowbar.

5. Detached (48.1%). The type of dwelling was detached (e.g., house, bungalow - chalet, bungalow or a farm building).

6. Dog (1.4%). A guard dog occupied the dwelling at the time of the burglary.

7. Enclosed (46.7%). The dwelling had an enclosed frontage or rear supporting unobservable access.

8. Exit prepared (0.5%). Exit point from the dwelling was prepared after entry to dwelling was gained.

9. Force (44.9%). Entry to dwelling was gained by using force (including bodily force or force with an instrument).

10. Gratuitous mess (5.7%). The offender 'ransacked' or made more than the necessary mess in the commission of the burglary (including defecation in dwelling, urination in dwelling, unnecessary damage caused inside dwelling and/or ransacking).

11. Intrusive search (34.3%). Search of dwelling was deemed intrusive when any of the following occurred: drawers or cupboards searched, drawers removed and/or food/drink consumed.

12. Jewelry stolen (36.7%). Jewelry was stolen, includes costume and/or valuable.

13. More than five items (27.6%). Offenders stole over five items.

14. Multiple search (41.9%). Searched most rooms (bedrooms, bathrooms, kitchen, lofts/roof space and lounge/dining room).

15. Nil stolen (9.5%). The offender did not take anything from the property.

16. No search (25.2%). The offender made no search of the dwelling. Either (a) property at specific sites within the dwelling were targeted (e.g., observable cash) or (b) only property from the immediate area surrounding the point of entry was taken.

17. Non-detached (48.1%). The type of dwelling was non-detached (e.g., terraced, flat or a high rise dwelling).

18. Poorly maintained (4.8%). The dwelling appeared to be poorly maintained by the occupants from the outside.

19. Private search (21.9%). The search of the dwelling targeted most bedrooms, bathrooms or the master bedroom only.

20. Secured (11.0%). Dwelling was secured by the offender to exclude occupants or observation by neighbors (e.g., closing curtains, wedging and locking internal doors, pacifying a dog with drugs or food, shutting a dog in a room, cutting the electricity supply to the dwelling and/or look out used).

21. Security light (9.0%). The dwelling was equipped with one or more security lights.

22. Sentimental (38.1%). Items stolen from the dwelling were deemed to have sentimental value to the owner (including antiques, clocks and watches, glass or crystal, gold jewelry, other jewelry, paintings and/or silverware).

23. Sign of occupancy –car (11.9%). A car was in the driveway of the dwelling indicating probable occupancy.

24. Sign of occupancy – curtains drawn (12.4%). Curtains of the dwelling were drawn indicating probable non-occupancy.

25. Sign of occupancy – light (9.0%). Lights in the dwelling were on indicating probable occupancy.

26. Special disposal (25.7%). The offender would likely have to make special arrangements in order to sell the items taken (including banking items and bank documents, china and porcelain, glass and crystal, paintings and/or silverware).

27. Untidy search (39.7%). Search of dwelling was conducted in an untidy fashion (e.g., drawers were removed and tipped out).

28. Well maintained (57.1%). The dwelling appeared to be well maintained by the occupants from the outside.

Appendix B

Content Dictionary for Serial Rapist Crime Scene Behaviours (Bennell et al., 2009; Melnyk, 2008)

Thirty-six variables were created from a content analysis of victim statements in order to provide a list of elements common to offences. All variables are dichotomous with values based on the presence (1) or absence (0) of each category of behaviour. The frequency for each behaviour is given in percentages following the variable name. A description of the categorization scheme in alphabetical order is given below.

1. Anal (14.3%). This variable refers to the offender penetrating or attempting to penetrate the victim's anus.
2. Apologize (7.1%). This variable refers to the offender using apologetic language directed at the victim at some point during the offence.
3. Ask (24.6%). This variable refers to the offender asking questions or being inquisitive about the victim.
4. Bind (20.6%). This variable refers to the use, at any time during the attack, of any article to bind the victim (excluding restraint by the offender's hands).
5. Blindfold (25.4%). This variable refers to the use, at any time during the attack, of any physical interference with the victim's ability to see (excluding verbal threats to the victim to close eyes or the use of the offender's hands).
6. Blitz (5.6%). This variable refers to the offender using a sudden and violent attack to overpower the victim; distinct from surprise attack because of the use of violence.

7. Compliment (10.3%). This variable refers to the offender complimenting the victim (e.g., on appearance).

8. Con (20.6%). This variable refers to the offender approaching the victim by giving a false impression of legitimacy (i.e., a false story, asking questions, etc.).

9. Cunnilingus (19.8%). This variable refers to the offender performing a sexual act on the victim's genitalia or attempting to perform such a sex act using his mouth.

10. Demand goods (25.4%). This variable refers to the offender approaching the victim with a demand for goods or money. This variable specifically relates to initial demands.

11. Demean (17.5%). This variable refers to the offender demeaning or insulting the victim (e.g., using profanities directed at the victim or women in general).

12. Disguise (15.9%). This variable refers to the offender wearing any form of disguise.

13. Display weapon (51.6%). This variable refers to the offender displaying a weapon in order to control the victim.

14. Extend time (14.3%). This variable refers to the offender extending the time spent with the victim after the actual attack.

15. Fellatio (24.6%). This variable refers to the offender forcing the victim to perform oral sex.

16. Force (victim participation) (38.9%). This variable refers to the offender forcing the victim to physically participate in the sexual aspects of the offence.

17. Force (victim sexual comment) (5.6%). This variable refers to the offender forcing the victim to make sexual comments.

18. Forensic awareness (23.0%). This variable refers to the offender showing knowledge of forensic procedures (e.g., fingerprints, DNA).

19. Gag (15.1%). This variable refers to the use, at any time during the attack, of any article to prevent the victim from making noise (excluding the temporary use of the offender's hand).

20. Identify (26.2%). This variable refers to the offender taking steps to obtain from the victim details that would identify her (e.g., by examining the victim's belongings).

21. Imply (9.5%). This variable refers to the offender implying that he knows the victim.

22. Kiss (37.3%). This variable refers to the offender kissing or attempting to kiss the victim.

23. Reassure (18.3%). This variable refers to the offender using reassuring or comforting language.

24. Reveal (26.2%). This variable refers to the offender revealing information about himself.

25. Sexual comment (52.4%). This variable refers to the offender making sexual comments during the attack.

26. Steal (identifiable) (7.9%). This variable refers to the offender stealing items from the victim that are recognizable as belonging to the victim (e.g., the victim's wallet).

27. Steal (personal) (10.3%). This variable refers to the offender stealing items from the victim that are personal to the victim, but not necessarily of any great value in terms of re-saleable goods (e.g., photographs or letters).

28. Steal (unidentifiable) (34.9%). This variable refers to the offender stealing items from the victim that are not recognizable as belonging to the victim (e.g., cash).

29. Surprise (92.1%). This variable refers to the offender using a method of approach consisting of an immediate attack on the victim.

30. Tear clothing (15.9%). This variable refers to the offender forcibly tearing the victim's clothing.

31. Threat (no report) (29.4%). This variable refers to the offender threatening the victim that she should not report the incident to the police or to any other person.

32. Threat (verbal) (11.9%). This variable refers to the offender threatening the victim using insults and profanity at some time during the attack (excluding threats not to report the incident).

33. Vaginal (front) (74.6%). This variable refers to the offender penetrating or attempting to penetrate the victim's vagina from the front.

34. Vaginal (rear) (17.5%). This variable refers to the offender penetrating or attempting to penetrate the victim's vagina from the rear.

35. Violence (multiple) (12.7%). This variable refers to the offender perpetrating multiple acts of violence against the victim (e.g., multiple punches).

36. Violence (single) (30.2%). This variable refers to the offender perpetrating a single act of violence against the victim (e.g., a single slap).